The Evolution of the Lactase Persistence Phenotype

Charlotte Mulcare

A thesis submitted for the Doctor of Philosophy degree
at the University of London

London 2005

The Centre for Genetic Anthropology
The Galton Laboratory
Department of Biology
University College London

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

The ability to digest significant quantities of the dissacharide lactose is dependent upon high expression of the enzyme lactase in the small intestine. Downregulation of lactase occurs in the vast majority of adult mammals, and most humans lose the ability to produce high levels of lactase after weaning (lactase non-persistence) whereas others keep high levels into adult life (lactase persistence). This variation in adult enzyme expression is under genetic control, and frequencies of the two phenotypes vary in throughout the world. A correlation between a longstanding culture of dairying and a high frequency of lactase persistent individuals in a population group has been reported. This has led to the theory that the high levels of lactase persistence seen in some groups could be the result of natural selection. This thesis examines variation in and around the lactase gene in to explore the possible role of natural selection in explaining modern frequencies of lactase persistence.

A series of single nucleotide polymorphisms (SNPs) in the vicinity of the lactase gene have alleles known to associate with lactase persistence in Northern Europe. These were investigated in a series of 4024 individuals from 73 population groups throughout the world to generate a global distribution that could be compared with anthropological data, and previously reported frequencies of lactase persistence. One of these SNPs, a C-T transition located –13.9kb upstream of the lactase gene, was reported during the course of this thesis, (Enattah et al 2002), with the T allele proposed as a putative cause of lactase persistence. The haplotypic background of this allele is reported here.

Although frequencies of the –13.9kb*T allele could account for reported frequencies of lactase persistence in European groups, this was not the case in sub-Saharan Africa, and also in some Eurasian groups. These data indicate either that –13.9kb*T is only closely associated with a true causative and, as yet, unknown mutation, or that there is heterogeneity of causes for the lactase persistence trait. Although –13.9kb*T was absent in East Africa, it did exist at low frequency in Cameroon, where an association with Fulani ancestry was

observed. Y-Chromosome data for this group supported a previously reported hypothesis for a Eurasian back-migration event in the Northern region of Cameroon, which might explain the presence of the –13.9kb*T allele in this region.

5 microsatellite loci in and around lactase gene were used to measure and compare intra-allelic diversity for 3 SNP haplotypes. A series of small families from 9 population groups were used to resolve compound SNP and microsatellite haplotypes. The compound SNP haplotype associated with lactase persistence was found at a disproportionately high frequency for its associated microsatellite diversity when compared to SNP haplotypes that do not associate with lactase persistence in Western Europe. This is interpreted as strong evidence for positive selection pressure favouring lactase persistent individuals in this region.

For my parents
with much love

Table of Contents

Contents of figures

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

Contents of Tables

Abbreviations and acronyms

A or a	adenine
AMOVA	Analysis of Molecular Variance
ASD	Average squared distance
BCE	Before Common Era (equivalent to BC, Before Christ)
BSA	Bovine Serum Albumin
bp(s)	base pair(s)
C or c	cytosine
°C	Degree(s) Celsius
CE	Common Era (as AD, Anno Domini)
CEPH	Centre d'Étude de Polymorphisme Humain
cM	centiMorgan
dbSNP	SNP database at NCBI
df	degrees of freedom
DMSO	Dimethyl sulfoxide
DNA	Deoxyribose Nucleic Acid
EDTA	ethylene diaminetetra acetic acid
EtOH	Ethyl alcohol (ethanol)
G or g	guanine
*Hinf*1	a restriction endonuclease derived from Haemophilus Influenzae
*Hinp*1	a restriction endonuclease derived from Haemophilus Influenzae
PCR	Polymerase Chain Reaction
LCT	Lactase
LD	Linkage Disequilibrium
M	Morgan
mA	milliamps
Mab	Mono-clonal antibody
Mb	Megabase
ML	Maximum Likelihood
$MgCl_2$	Magnesium Chloride
mRNA	messenger Ribose Nucleic Acid
MspI	a restriction endonuclease derived from Moraxella species
MSAT	Microsatellite tandem repeat sequence
N/A	Not applicable
NaCl	Sodium chloride
NlaIII	a restriction endonuclease derived from Neisseria lactamica
No.	Number
Oligo	oligomer
OMIM	Online Mendelian Inheritance in Man
p	probability
pH	Potential of hydrogen
RFLP	Restriction Fragment Length Polymorphism
rpm	Revolutions per minute
RNA	Ribose Nucleic Acid
SDS	Sodium diodecyl sulphate
SNP	Single Nucleotide Polymorphism

SSCP	single strand conformation analysis
STR	short tandem repeat
T or t	thymine
TAE	Tris-acetate EDTA
TBE	Tris-borate EDTA
Taq	Thermus aquaticus (DNA polymerase)
TCGA	The Centre for Genetic Anthropology
μ	Micro (prefix)
UV	Ultra-violet
V or v	volt
χ^2	*Chi* square statistic

Acknowledgements

One thing that has become clear to me over the past three years is the importance of collaboration in science, and since I have worked between two laboratories and two groups, the Galton Laboratory and the TCGA laboratory, the list of people to be thanked is quite substantial!

First and foremost, this thesis would not have been possible without the continued guidance, wisdom and endless patience of my supervisors, Mark Thomas and Dallas Swallow. Both have shared their extensive experience and skill with great generosity, and it has been a privilege and pleasure to work with them. In particular, both made thoughtful and detailed comments on the final version of this thesis.

Especial thanks also go to Lynne Vinal and Abigail Jones who, as lab managers, provided much time and advice in teaching me many of the basic lab skills I have relied on over the past four years.

Dr. Mike Weale wrote the statistical procedure that was used for comparing the frequency of reported lactase persistence with observed genotype *(GenoPheno)* and has also been amenable to pestering for statistical advice. Dr. Michel Stumpf kindly provided his 'Syssiphos' program for identifying selection signatures, and gave time to answer questions regarding its operation.

I was funded by a BBSRC CASE studentship, and the BBSRC provided the financial backing for this research. Dr. Neil Bradman has, as my industrial sponsor, provided financial support during the PhD, but has also made a valuable academic contribution with regard to conceptualising the work on the distribution of –13.9kb*T allele in Africa. I also thank Dr. Roger Blench and Dr. Clare Holden for their time and helpful discussion relating also to that area of research.

Dr. Katri Peuhkuri, Dr. Kasja Kajanda and Dr. Riita Korpela collaborated with our group when investigating the Finnish data set, and had previously undertaken the phenotyping for the collection of Finnish samples in Helsinki. Ms. Lupe Polanco first trialled the C-13.9kbT typing protocol on the Finnish and UK cohort panel referred to in the Poulter et al (2003) paper and in chapter 3 of the thesis. I also thank the PhD students who worked with Dallas before me, Dr. Ed Hollox and Dr. Clare Harvey, for their theses, which provided the foundation of knowledge that this research has built on. Ms. Kate Ingram, who has recently embarked on a PhD in lactase kindly helped with some final checks for me, and also has provided her thoughts and good company.

Two undergraduate students worked with me on their final year projects on lactase. In the Darwin Building, Alex Murray was supervised by me and worked on a project, which involved extracting the samples from the Extreme Northern Region of Cameroon used in chapter 4, and she also completed most of the practical work of typing them. In Wolfson House, I worked with Fang Wei Lin on her final year project, and she similarly made a substantial contribution by typing the East Asian and primate samples described in chapters 3 and 7.

The Centre for Genetic Anthropology has an extensive collection of DNA, which has been collected and extracted over the years by a significant number of its members, partners, and undergraduates. These include: Dominic Gormis, Esther William, Tanelli Helenius, Jim Wilson, Pieta Nasanen, John Greenhalgh, Carl Gierstorfer, Jane Moore, Richard Phillips, Katya Bulgina, Ali Barwhani and Noreen von Cramon-Taubadel, all of whom collected and extracted and/or tested for Y chromosome many of the African and Eurasian DNA samples used from the TCGA collection.

In particular, I also thank Dr. Mark Thomas, Dr. Neil Bradman and Mr. Krishna Veeramah for allowing me to use the unpublished TCGA Y-Chromosome data for the observations and table in chapter 4.

From the Galton laboratory, samples were kindly donated by the following collaborators: Dr.Vladimir Ferak and Mr. Marek Svalik provided the Roma samples; Professor Trevor Jenkins and Dr.Amanda Krause provided the San and Bantu samples; Dr. Nilmani Saha provided the Indian, Malay and Chinese samples and Dr. Geoff Daniels provided the Japanese samples.

The TCGA lab and the Galton lab both host many other researchers and workers whose company has been a great pleasure these last few years, and who have given invaluable advice ranging from the coaxing of photocopying machines to PC and MAC compatibility dilemmas, and, as importantly, have been a pleasure to share my breaks with. They are (in alphabetical order): Mr. Ed Almond, Dr. Ian Barnes, Dr. Cathy Bridge, Dr. Liz/Beth Caldwell, Mr. Lewis Dartnell, Dr. Adil Elamin, Mr. Ian Evans, Dr. Ben Fletcher, Mr. Andrew Loh, Dr. Karine Rousseau, Mr. Dave Turner, Mr. Krishna Veeramah (who kindly loaned me his PC and statstically-adroit brain on many occasions), and Mr. Lorenzo Zanette. Ranji Arasaretnam has kept the lab in working order, and kindly poured some gels for me for some of my final checks. Ms. Ana Teixeira has been a wonderful office partner, has also loaned me her PC and has taught me the true art of coffee making, Portuguese style. Dr. Sarah Joshua suffered the burdenous responsibility of reading parts of my first draft, and very kindly provided a second pair of eyes for data checking gels.

Outside of the world of genetics, several people bear responsibility for keeping me as close to sanity as I was ever likely to get in the course of this PhD. This thesis is dedicated with much love to my parents, who have provided their ongoing support in every possible way. , I also acknowledge with gratitude their persistent and vocal belief that this is the best thesis on lactase persistence that they will ever read.

Thanks also to Patrick, Zoe, Ben and Clara Mulcare, and to Marion, Geoff, Richard and Robert Ward, who have provided moral support and access to playstation at crucial moments. Ms Mellie Naydenova has especially helped me keep the voices at bay (albeit at the cost of my liver function), as have Ms Celina Cundy, Ms Elisabet Gunzi, Dr. Caroline Horne, Dr. Lucy Jessop, Ms Ashlie Johnson, Dr. Priya Patel, Mr. Nikin Patel, Mr Craig Richardson, Ms Marie Richardson, Dr. Shivangi Thakore, Ms Caroline Lang, Ms Sandra Martelli, and Mr Will Niblett with their most excellent company and friendship.

Last but certainly not least, I would like to thank my long-suffering partner, Dr. Marios Costambeys. He has endured with great fortitude my descriptions of PCR, continually putting up with the changing fortunes and directions of the thesis, and has provided the continuous support that I have relied on so heavily over these last years. May Chelsea ever win at Stamford Bridge.

Chapter One

Introduction

'With regard to milk, it should not be given to all, but only to those who digest it well and perceive no symptom ... somebody can be sick all the time, no matter in what way he prepares it (milk) and someone else, trying to use the milk, had no trouble, for he digested it well and had no hyperacidity, or eruction, or gas'

'De sanitate tuenda' Galen, 2nd Century AD
[Trans. Robert Montraville Green 1951]

The observation that not all humans can digest fresh milk is far from new; in Ancient Rome, the medical notebook of Galen records his observation that some healthy adults experience a series of symptoms after drinking milk, symptoms currently known as 'lactose intolerance', whereas others are able to drink milk without complication. In the past few decades, the nature of this milk drinking ability has been extensively investigated, and has attracted interest from evolutionary geneticists.

More recently, two phenotypes have been reported. Downregulation of lactase occurs in the vast majority of adult mammals, and most humans lose the ability to produce high levels of lactase after weaning (lactase non-persistence) whereas others keep high levels into adult life (lactase persistence).

Since the majority of early studies were performed on Northern Europeans where high levels of enzyme expression in adulthood is common, it was originally assumed that this was the 'normal' condition, and that reduced expression was pathological, hence the early naming of the phenotype 'lactase deficiency'. Other frequently used terms include 'lactase restriction', 'hypolactasia', 'lactose digestion' and 'maldigestion' (for summary, see Sahi et al 1974). Later studies showed variation in frequency of the phenotypes between populations and it became apparent that, world-wide, low levels of lactase in human adults was more common than high levels (for review, see Swallow 2003), and so should not be considered a pathological trait. For this reason, many authors have now adopted the

terms used in this thesis, 'lactase persistence' and 'lactase non-persistence' to reflect the fact that although two phenotypes exist in human populations, both should be considered normal states.

The key calorific component of fresh milk is the sugar lactose, which is hydrolysed by the enzyme lactase phlorizin hydrolase (lactase). Although all healthy children can digest fresh milk, only some adults express lactase at sufficient levels to digest high levels of lactose, and the frequency of such individuals varies significantly between global population groups. The level of expression of the enzyme has been shown to be related to the level of transcription and, although a causal polymorphism has not yet been conclusively identified, some alleles show a strong association with lactase persistence (for recent review, see Swallow 2003). This project studies variation in and around the human lactase gene in different populations as a tool to investigating possible historical causes for the modern distribution of lactase persistent phenotypes.

Information and techniques from different disciplines were used in this project; consequently, this introduction is structured to provide an overview of four areas of research. The first section focuses specifically on lactase itself and will describe the molecular aspects of the study, from the characterisation of the protein to its expression, and the role of lactase in digestion. The evidence for a genetic control of variable lactase expression in different populations is considered, and studies demonstrating association between genetic variants and the two phenotypes are reviewed. The second section summarises the global frequencies of the lactase phenotypes, current theories explaining their distribution, and relevant clinical aspects, including the various methods of diagnosis and their reliability. The third section discusses theories about the development of pastoralism during the Neolithic transition, the history and nutritional impact of milk drinking and the possibility of dietary adaptation. The final section introduces the techniques that will be used in the thesis, specifically the methods used in population genetics for identifying signatures of selection, migration and drift.

Part 1

Molecular Aspects of Lactase Persistence

1.1.1 The digestion of lactose

1.1.1.1 The role of lactase in digestion

As early as 1903, Rohman and Nagano suggested that the inability to digest fresh milk was due to an inability to digest high levels of lactose, the key carbohydrate in fresh milk. Lactose is a disaccharide composed of glucose and galactose that must be hydrolysed to its monosaccharide components before absorption in the small intestine. This hydrolysis reaction is catalysed by the enzyme lactase-phlorizin hydrolase (lactase). It was subsequently shown that variation in the ability to digest large volumes of lactose was dependent on the expression of the lactase enzyme, in some human adults, lactase was not present at sufficient levels to facilitate digestion of the substantial quantities of lactose found in fresh milk (Kern and Struthgers 1966).

1.1.1.2 The site of lactose digestion in the gut

Lactose, most usually found in milk, is ingested through the mouth and travels down the oesophagus to the stomach. Here, it is mixed with gastric juices secreted by the stomach lining to form chyme, which moves from the stomach via the pylorus and ileococcal valve into the small intestine (comprising the duodenum, jejunum and ileum) and through peristalsis to the large intestine, comprising the caecum, colon and anus (for example, Guyton and Hall 1996).

A series of early studies suggested that the most active site of lactose hydrolysis was the small intestinal mucosa (for example, Dahlqvist and Borgstrom 1961) and a later study by Newcomer and McGill (1966) showed that, correspondingly, lactase expression levels were highest in these areas. Their study investigated the longitudinal distribution of disaccharidase activity throughout the small intestine, and showed that lactase activity is first observed in the duodenum, reaches its highest levels in the jejunum and declines thereafter.

The epithelium of the small intestine is comprised of villi that increase the surface area over which digestive reactions can occur. The depressions between two adjacent villi are known as crypts, and these contain stem cells that are able to divide and produce daughter cells that differentiate. These daughter cells migrate along each villus to replace epithelial cells that are continually lost from the tips of the villi. The enterocyte cells that form the major part of the epithelial lining of the small intestine increase the absorptive surface area through their brush-border membrane composed of micro-villi. Mucous cells that are also present in the epithelium secrete protective mucus in the intestine (Guyton and Hall 1996). Lactose is located in the micro-villus membrane on the enterocytes and expression of lactase is greatest at the midpoint between the apex or pinnacle of the villi and the crypts (Skovbjerg et al 1982). Figure 1.1[1] below shows a histological tissue sample taken from a human jejunal biopsy, using immunohistology to detect lactase expression.

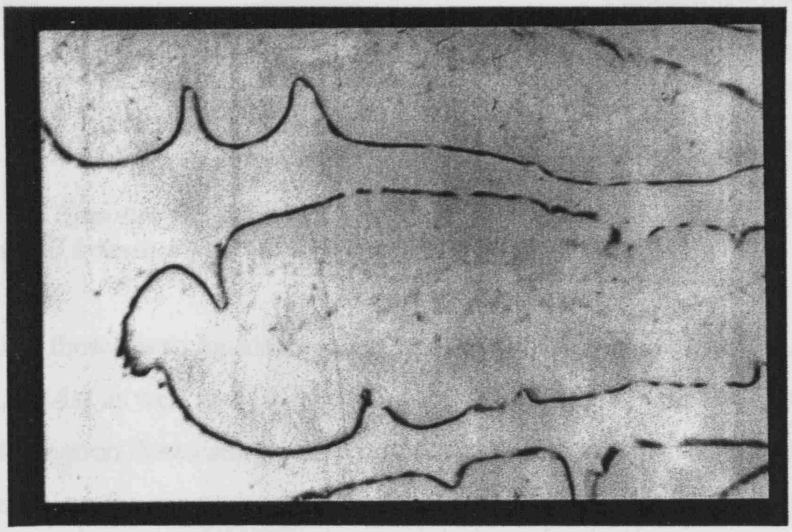


Fig 1.1 A histology sample from the jejunum showing a section of villi for the lactase enzyme immunostained using the monoclonal antibody mlac 4. *The black outline shows where lactase is present on the villus and holes in the outline show mucus producing goblet cells.*

[1] Photograph kindly provided by Professor Dallas Swallow

1.1.1.3 Comparison of disaccharidases

Lactase is one of a series of enzymes found in the small intestine that hydrolyse disaccharides, the other key molecules being sucrase-isomaltase, maltase-glucoamylase and trehalase (see table 1.1). Genetically determined reduction of these other enzymes also occurs, although is thought to be rare in most populations.

Enzyme	Metabolites	Observed variation in expression?	Other comments
Sucrase isomaltase (*SI*)	*Sucrose* → Fructose, Glucose *Isomaltase* → Maltose, Glucose	Deficiency found in 10-15% of Inuit (for example, Gudmand-Hoyer 1987); lower sucrase activity observed in Bantu groups (Veitch et al 1998)	Amino acid changes leading to deficiency have been reported in a few rare European cases (for example, Ouvendijk et al 1996).
Maltase glucoamylase (*MGAM*)	*Maltose, Amylose* → Glucose	None so far shown to have a genetic basis	A combined *SI/MGAM/LCT* deficiency has been reported (Nichols et al 2002).
Trehalase	*Trehelose* → Glucose	Deficiency is an autosomal recessive trait found at 8% frequency in Greenland (McNair et al 1972)	Trehelase deficient individuals might have problems if trehalose-rich fungi were ingested; trehelose is now more commonly found in artificial sweetners.

Table 1.1 Summary of different disaccharidases involved in digestion and found in the small intestine

Several of these disaccharidases show evidence of internal duplication, specifically, *SI* and *MGAM* as well as *LCT* (see also section 1.1.2.1), and possibly arose from gene duplication. Interestingly, the amylase gene family shows evidence of polymorphism for gene copy in adult humans, which affects the level of enzyme activity (Groot et al 1989). However, the high frequency in human populations of both lactase phenotypes, in comparison with others, remains distinctive.

1.1.2 Structure and expression of the Lactase enzyme

1.1.2.1 Properties of the lactase enzyme

As its full name suggests, lactase also hydrolyses phlorizin, a sugar found in the tips and roots of Rose plants and some forms of seaweed, as well as a number of other glycosides and galactosides, such as cellulobiose, cellotriose, cellotetrose and to a lesser extent, cellulose (for review, see Troelsen 2004). Lactase has two active sites, a ß-galactosidase site that hydrolyses lactose and other ß galactosides, and also a ß-glucosidase site that hydrolyses phlorizin, flavonoid glucosides and pyridoxine-5'-ß-D glucoside (Nemeth et al 2003, Mackey et al 2002).

Sequence analysis shows four internal repeat sequences, suggesting that the lactase gene has been subject to two gene duplication events in its evolution with four domains containing potential active sites (Troelsen 2004). A precursor to the enzyme with these four domains and a molecular weight of 245,000 undergoes glycosylation and proteolytic cleavage to attain its mature active state (Naim et al 1995). The mature enzyme has only two domains and a molecular weight of 160,000, and the remaining polypeptide from the cleavage is likely to have a role as a chaperone, though it has no disaccharidase function (Naim et al 1994, Oberhalzer et al 1993). Following intra-cellular transport, the mature enzyme is anchored to the cellular membrane at the C-terminus, with the majority of the protein being extra-cellular (Troelsen 2004).

1.1.2.2 Expression of the lactase enzyme in humans

In humans, lactase is first expressed during fetal development in the second trimester of pregnancy and levels increase in the third trimester (Doell and Kretchmer 1962). The peak of enzyme production usually occurs in the neonatal phase of life (Auricchio et al 1965, Dahlqvist and Lindberg 1966, Antonowiez and Lebenthal 1977), and, in most humans as with most other mammals, declines to levels of between 5 - 20% of neonatal levels after weaning (Blaxter 1961).

However, in some population groups, notably those with a history of pastoralism, lactase levels remain high throughout adulthood (Holden and Mace 1997).

1.1.2.3 Expression of the lactase enzyme in other mammals

Since the Oxford English Dictionary definition of a mammal includes *'having mammary glands that secrete milk to feed the young'* (OED 2005 online[2]), and since most mammalian milk contains lactase, it is perhaps unsurprising that investigations of lactase activity across mammalian species indicates high expression in the small intestine. There are some exceptions, such as sea lions and other Pinnepedia, where lactose is not excreted in milk, and correspondingly, the neonates do not show evidence of lactase activity (Crisp et al 1987). Most mammals appear to show downregulation of lactase after weaning, for example, in sheep (Lacey et al 1994), rabbits (Sebastio et al 1989, Villa et al 1993), rats (Sebastio et al 1989, Buller et al 1990, Lee et al 2002) and pigs (Troelsen et al 2003, Pie et al 2004).

There are few studies on non human primates and it has been reported that the baboon (Welsh et al 1974) and the 'galago monkey' (Wen et al 1973) show a decline in lactase after weaning, but a study on macaques (Wen et al 1973) showed 9/10 expressed high levels of lactase as adults, which may even suggest heterogeneity of lactase persistence in other species as well as humans. However, this may also be because delayed weaning in many primates prolongs lactase expression as part of the life cycle rather than a genetically controlled mechanism (for example, see Harvey et al 1987). No studies have been undertaken on higher primates and so it is possible that investigation into lactase persistence status of non-human great apes might reveal evidence of either or both phenotypes.

[2] Website reference: http://dictionary.oed.com/

1.1.3 Is the ability to digest fresh milk a genetically controlled trait?

Two possible explanations were initially suggested when variation in the amount of lactase expression between human adults was observed: first, that a genetic mutation enabled some individuals to continue producing the enzyme (for example, Howland 1921) or, second, that the expression of lactase was controlled by non-genetic factors, such that continuous milk drinking throughout adulthood could stimulate higher levels of lactase (for example, Bolin et al 1969). This debate is of significance since a prerequisite for using genetics to investigate the evolutionary processes responsible for the modern distribution of the lactase persistence phenotype depends on lactase persistence being a heritable trait.

1.1.3.1 Early studies

In 1906, a study by Plimmer had suggested that the lactase enzyme could not be induced through increased consumption of lactose by adult rabbits and rats, but it was not until the 1960s and 1970s that this experiment was repeated using a group of adult humans. Again, it was demonstrated that the enzyme did not increase through dietary stimulus (for example, Cuatrecasus et al 1965, Keusch et al 1969). From the non-genetic perspective, one possible explanation for this observation is that, once down-regulated, the high lactase levels of infancy cannot be restored, but if an individual continues to drink milk throughout childhood, adolescence and into their adult life, expression of lactase can be maintained at high levels. However, this was shown not to be the case in studies in which milk was fed continuously to children who nevertheless showed evidence of decline of lactase (Flatz 1977). Also, some adults who did not habitually consume fresh milk were shown to digest lactose without the need of any period of time for adaptation (Flatz and Rotthauwe 1971).

1.1.3.2 Family studies

The first strong indications that the genetic hypothesis was correct came from early family studies, though many of these had small sample sizes, and in some cases phenotypic information was incomplete for a given family. Sahi et al (1973)

published the most comprehensive family study, in which 327 Finnish individuals from a series of families were tested for lactase persistence status. All participants were over the age of twenty, to minimise the potential confounding effect of cases of delayed onset of lactase non-persistence. Families were placed into three groups depending on their parents' phenotype. In one group, both parents were digesters, in one group, neither was, and in the last group, only one parent was able to digest lactose. The inheritance of lactase phenotypes conformed to an autosomal Mendelian inheritance pattern, with non-persistence as the recessive condition. Later family studies in Mexico (Lisker et al 1975), Nigeria (Ransome-Kuti et al 1975) and amongst Native American families (Newcomer et al 1977) showed the same pattern of Mendelian inheritance in different populations, strongly supporting the genetic hypothesis.

1.1.3.3 Twin studies

Twin studies are frequently used to distinguish genetic effects (which should affect monozygotic twins identically but not dizygotic twins) and environmental effects (which should affect siblings similarly). A study on 102 adult twin pairs from Hungary using a variety of indirect phenotyping methods showed that the lactase persistence phenotype was concordant in monozygotic but not dizygotic twins (Metneki et al 1984). Metkneki concluded that his evidence supported the genetic hypothesis.

1.1.3.4 Intestinal Lactase activity

A study on autopsy samples taken from the small intestines of 75 English individuals aged between 11 and 88 years performed enzyme assays on both sucrase and lactase (Ho et al 1982). Only individuals without protracted illnesses or evidence of digestive disease were used. A wide range of lactase activities were observed; however, when the ratio of sucrase/lactase was investigated, a trimodal distribution could be observed, consistent with two homozygous and heterozygous levels of expression. This trimodality could not be observed in lactase levels alone, since there was (as might be expected) some variation in the quality of the sample.

Similarly, a trimodal distribution was discovered in a series of German individuals' maltase/lactase ratios (Flatz 1984).

Although in both cases heterozygote individuals could be distinguished from homozygote persistent individuals, activity in heterozygotes is sufficient to digest the lactose in a lactose tolerance test (see section 1.2.2), and so lactase persistence is considered a dominant trait. However, it is perhaps likely that heterozygotes are more prone to suffer secondary hypolactasia (the loss of high levels of lactase expression due to certain illnesses) than homozygote persistent individuals. This is particularly relevant since, in persistent individuals, lactase is one of the last enzymes to recover high levels of expression following gastro-intestinal illness (for example, Phillips et al 1988).

1.1.3.5 What was the ancestral trait?

It is now commonly agreed that lactase persistence is a variable trait in modern humans, which is under genetic control (MIM 223100). The next consideration is which of the two phenotypes is ancestral, since this affects how models of the distribution of both phenotypes are weighted. There are two main strands of evidence to suggest that lactase non-persistence is the ancestral state. As discussed, in most other mammals, lactase levels decrease substantially after weaning, which suggests that from an evolutionary perspective, enzymatic decline is the usual mammalian state. Also, there is a higher prevalence and wider distribution of the lactase non-persistence phenotype in modern humans (for example, Swallow 2003). However, the possibility always remains (albeit unlikely) that the early hominids were lactase persistent.

1.1.4 The Lactase Gene

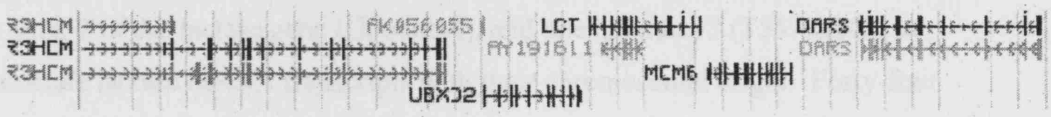
1.1.4.1 The cloning of the lactase gene

In 1988, the primary structures of human and rabbit lactase genes were deduced from cDNA sequences. The human primary translation products suggested that there were 1927 amino acid (Mantei et al 1988). In the same year, the gene was

mapped to a gene rich region on chromosome 2 (Kruse et al 1988). Once a genomic clone was available, further characterisation of the gene was undertaken, and it was shown that lactase has 17 exons spanning a distance of approximately 70kb (Boll et al 1991). This information, and an abbreviated formal nomenclature for lactase [*LCT*-MIM 603202] has been utilised since 1991 for more investigation of lactase; since complete sequencing of the human genome, it has become evident that the actual size of lactase is somewhat smaller than 70kb, approximately 50kb[3]. Boll et al's study also compared the coding sequences from 11 individuals, and identified 14 variants, none of which showed complete association with the lactase persistent phenotype. However, three of these polymorphisms involved non-synonymous substitutions. These were: valine to isoleucine (G666A), alanine to threonine (G3297A) and methionine to asparagine (G4927A).

1.1.4.2 Localisation of the Lactase Gene

Harvey et al (1993) used fluorescence in-situ hybridisation to further resolve the location of the gene, which was localised to 2q21, confirming earlier work using the Centre d'Étude du Polymorphism Humain (CEPH) family linkage maps (NIH/CEPH Collaborative mapping group 1992). Further genomic cloning of the flanking regions characterised a gene 3kb directly upstream of the transcription initiation site of lactase named *MCM6* (Harvey et al 1996). This is a homologue of a yeast gene involved in cell cycle control. Further upstream is the DARS gene (Escalante and Yang 1993), and downstream is the *UBXD2* gene (alias KIAA0242) and *R3HDMM* (alias KIAA 0029). The diagram below, taken from the human genome browser, shows the genes immediately surrounding lactase (*LCT*).



[3] As shown, for example, on the human genome browser http://genome.ucsc.edu/cgi-bin/hgGateway

Fig 1.2 A schematic diagram to show the relative positions and sizes of genes spanning a region of approximately 550kb surrounding the lactase (LCT) gene. *The arrows indicate the direction of transcription and spaces indicate the relative size of the inter-gene regions*

1.1.5 mRNA levels and lactase expression

1.1.5.1 Varying levels of lactase expression

Most studies using RTPCR, and other methods, indicated that lactase mRNA was downregulated and that the mRNA levels correlate well with lactase expression (for review, see Troelsen 2004). A study by Maiuri et al (1994) showed that, in the jejunal tissue of non-persistent individuals, lactase expression was patchy, and that the extent of downregulation was not uniform, suggesting that under some circumstances that cells escape the downregulation mechanism.

1.1.5.2 The causal element of lactase persistence: cis- or trans-acting to the lactase gene?

The study of Ho et al (1982) and also Flatz et al (1984) that described a trimodal distribution of lactase activity inferred that the genetically determined element controlling differences in lactase levels was more likely to be cis-acting to and probably regulatory of the gene. The promotor and exons characterised by Boll et al (1991) and Lloyd et al (1992) were screened for polymorphisms that might associate completely with lactase persistence but, as none were found, it also remained possible that the causal element might be located some distance from the lactase gene.

A study by Wang et al (1995) used two previously identified polymorphisms, one in exon 2 of the lactase gene (G666kbA), and one in exon 17 (T5579kbC), to associate lactase mRNA transcripts with their chromosomal origin. Forty-four adult duodenal biopsies from London were used both to diagnose persistence status (using sucrase/lactase ratios) and to extract mRNA. If the causal element of lactase persistence were trans-acting, lactase persistent heterozygotes should show

approximately equal levels of mRNA transcripts from each chromosome. However, if the causal element were cis-acting, the expression would be independent, and one transcript would be higher than the other in heterozygous individuals. The study showed that in most individuals who were heterozygous for the exonic SNP, one allele of the gene was expressed at elevated levels compared to the other, and this correlated (on average) with intermediate levels of lactase activity, strongly supporting the cis-acting hypothesis.

*1.1.6 **Polymorphisms in and around the lactase gene***

1.1.6.1 Identification of a series of 'core' lactase haplotypes

Harvey et al (1995) further characterised variation in and around the lactase gene to determine the extent to which the polymorphisms were associated with each other. Portions of the lactase gene from a series of over 60 CEPH families were taken, and a combination of single strand conformation analysis (SSCA) and denaturing gradient gel electrophoresis (DGGE) was used to detect polymorphisms. Seven polymorphisms, including some described by Boll et al (1991) and Lloyd et al (1992) were tested. Haplotypes were then inferred, and only three common haplotypes, A, B and C, were identified (Harvey et al 1995) in a region of linkage disequilibrium extending around the lactase gene, encompassing a region of approximately 70kb (see Appendix B1).

Hollox et al (1999) found further variation in a region 974bp to 852bp upstream of the lactase gene. Although none of these observed variants showed a complete association with the lactase persistence phenotype, one polymorphism, a C-958T transition, was shown to affect protein binding. Hollox et al proposed that this might affect the timing of downregulation in children (1999), which is known to be variable (Wang et al 1998).

1.1.6.2 Distribution of haplotypes and the association of the 'A' haplotype with lactase persistence

The frequencies of the described core lactase haplotypes in a group of Europeans and Indians were investigated (Harvey et al 1998). Although A, B and C were common throughout Europe, the A haplotype was observed most frequently in Northern Europeans (sample groups from London), whereas B and C were more common in Southern Europe (sample group mainly from Italy) and India (Harvey et al 1998). These three haplotypes, and an additional E haplotype, were found in all populations, with a D haplotype observed in Europe and a G in India. The paper also showed data from 11 unrelated lactase persistent individuals, and it was hypothesised that the mutation causing lactase persistence was found on the background of the A haplotype. This theory was supported by the observation that the A haplotype was prevalent at high frequencies only in Northern Europeans, where the highest frequencies of lactase persistence are found.

Further polymorphisms were later identified, such that haplotypes could be inferred from 1338 chromosomes using 11 different markers ranging from –958kb upstream of the start of lactase, to 6236/7kb downstream from the start of the *LCT* gene. From these, 40 haplotypes were identified, although A, B and C remained the most common (Hollox 2000, Hollox et al 2001). The 'A' haplotype was again found to be most prevalent in Northern Europe, and followed a cline further east, though it was also found in Asia, parts of Africa and Indonesia. A further 'U' haplotype, common in Africa, was not frequently observed in Eurasia (Hollox et al 2001). Appendix B1 shows the allelic states of these core lactase haplotypes.

A greater diversity of haplotypes was observed in sub-Saharan Africa than in non-Africans (Hollox et al 2001), and the data also showed a greater frequency of the 'K' haplotype in sub-Saharan Africa, which was speculated to be the ancestral haplotype. This was on the basis that a panel of chimpanzee samples were characterised as being closest to the 'K' haplotype, which has ancestral states at the greatest number of SNP loci (Hollox 2000). Hollox et al (2001) proposed that the pattern of variation in Northern Europe, with such a high frequency of the A Haplotype, is the result of recent directional selection in Northern Europeans.

1.1.6.3 Investigating a cause for lactase persistence

Despite the evidence that a mutation in the vicinity of *LCT* was causative of lactase persistence and the extensive characterisation of polymorphisms in and around the lactase gene, identifying a causal mutation has proved difficult. The most probable mechanism is one affecting the expression of the gene at the level of transcription since, as discussed, reduced levels of mRNA has been shown in non-persistent adults. After this project was initiated, two SNPs in *MCM6* were shown to be highly associated with lactase phenotype in Finland (Enattah et al 2002). One, a C – T transition located –13.9kb upstream from the start of the lactase gene, showed 100% association with lactase persistence in the Finnish population, and the other, a G-A transition located –22kb upstream from the start of the lactase gene, showed association in 229 of 236 cases of persistent individuals.

These two SNPs were identified after an initial investigation of linkage and of microsatellite haplotype diversity and linkage disequilibrium analysis in 9 Finnish families. In these families, phenotype was known and phase was established using family pedigrees. Initially, a series of microsatellites spanning a region of 222.5kb, including *LCT,* were used to determine the extent of LD. Microsatellite haplotype analysis across 150kb identified a conserved region of 47kb (36kb of which comprises the *MCM6* gene) that was identical in lactase persistent alleles. Complete sequencing of this 47kb region identified 52 polymorphisms, two of which were extremely tightly associated with lactase persistence phenotype. One hundred and ninety-six 196 intestinal biopsies were used to determine persistence status in a population of Finnish individuals, and phenotype corresponded completely with genotype for the C-13.9kbT polymorphism, and, to a lesser extent, with the G-22kbA polymorphisms (Enattah et al 2002).

Forty non-Finnish individuals, (23 South Korean, 9 Italian and 8 German) who were all diagnosed as non-persistent were all homozygous for the –13.9kb*C allele. 938 anonymous Finnish blood donors were typed for the C-13.9kbT polymorphism

to determine frequencies of the genotypes, which corresponded to published frequencies of lactase persistence in the Finnish population. The distribution of the –13.9kb*C and –13.9kb*T alleles was also determined in a French population (n = 17) and also two American population groups, (n=92 European descent, n=96 African descent). The authors suggested that the frequencies of the genotypes corresponded with published data on persistence for those groups (Enattah et al 2002).

On the strength of these data, the –13.9kb*T allele has been proposed as a possible candidate for the cause of the lactase persistence phenotype. If the causal mutation has been found, two major repercussions are evident: first, the lactase persistence phenotype can be easily diagnosed using a DNA test, and secondly, extensive population studies using the allele as a proxy for persistence are possible. To determine whether the –13.9kb*T allele is truly causative of the trait, further investigation both within and outside Europe is necessary, and this ongoing work will be discussed in depth during this thesis.

PART 2

The Lactase Persistence Phenotype

1.2.1 The Lactase Persistence polymorphism

1.2.1.1 Defining the two phenotypes

In this thesis, the phenotype 'lactase persistence' refers to the continuing production of the lactase enzyme at the high levels produced in infancy, and 'lactase non-persistence' refers to individuals who downregulate production of lactase. The terms 'lactose digestor' and 'lactose non-digestor' are used in the context of diagnosing these two phenotypes, as will be discussed in section 1.2.2

A curiosity that has yet to be fully explained is the difference in age of onset of the decline of levels of enzyme expression between populations, since a significant range of ages has been observed. The latest age of onset is 20, observed as the upper threshold of the range in Finland, (Sahi and Launiala 1978, Sahi et al 1983) whereas in China the average is three years (Tadesse et al 1992). A study by Wang and colleagues (1998) showed that downregulation could be observed in the UK in a series of children from a gastro-intestinal clinic from the second year of life. It is possible that here some environmental factors are relevant to age of onset, such as infant viral disease, hormonal factors or even another genetically controlled mechanism. Whatever the reason for this variation, it is relevant to lactase research for two reasons: first, individuals under a certain age cannot be phenotyped with confidence because they may yet lose their lactase expression, and the 'cut off' age for which phenotype data can be accepted might vary between population groups. Secondly, from the population genetics perspective, individuals who become lactase non-persistent in their early twenties may have already had some selective advantage prior to the downregulation of lactase, and this might affect distribution of genotypes and the rate of increase in the frequency of the lactase persistent allele under selection.

1.2.2 Clinical implications and diagnosis of the two phenotypes

1.2.2.1 Clinical definitions

Although this thesis is concerned with the genetically determined phenotypes described above, the medical profession also recognises a series of conditions that may result in lactose malabsorption. 'Primary adult hypolactasia' is a clinical term that refers to the non-pathological reduction of lactase levels in adulthood, synonymous with the lactase non-persistence phenotype described here. Secondary hypolactasia refers to the loss of the enzyme lactase as a result of other causes, for example, through an illness such as Crohn's disease or coeliac disease (Hollox and Swallow 2002). There is also a very rare condition, 'congenital lactase deficiency', which is identifiable from birth, in which an individual does not produce lactase even in infancy. The locus for this condition has been assigned to 2q21, and claimed to be near to but separate from the lactase gene (Jarvela et al 1998).

1.2.2.2 Clinical Diagnosis of primary hypolactasia

In the absence of a definitive causal mutation, a test for the lactase persistence phenotype that can predict the likely levels of enzyme expression is of great importance in diagnosing the two phenotypes. As discussed, secondary hypolactasia also affects lactase expression, and so patients or volunteers must have no intestinal disease that might cause depleted lactase levels for pathological rather than genetic reasons (specifically, secondary hypolactasia as opposed to primary hypolactasia).

There are several different ways to establish phenotype. Biopsies of the jejunum or duodenum may be taken and enzyme activity measured directly. However, this is an invasive procedure, and so three indirect procedures are more commonly used: the blood glucose test, the urine galactose test and the hydrogen breath test. All three involve the patient or volunteer ingesting a lactose load, usually ~50 grams, and, if the lactose is digested, a rise in its constituent monosaccharides glucose and galactose can be expected. Levels of glucose are

recorded from a blood sample taken before and after ingestion, and similarly levels of galactose are measured using urine samples. The hydrogen breath test relies on the fact that hydrogen is produced during the fermentation of undigested lactose and can be measured in the breath (for review, see Hollox and Swallow 2002). It is conventional in the medical literature to describe lactose absorbers as 'negative' and non-absorbers as 'positive', that is, giving a positive result in a lactose tolerance test.

One problem with investigating the distribution of phenotype is the inaccuracy of the indirect tests, which do not show a complete correlation with assays of intestinal lactase (for example, Neale et al 1968). Several studies compare the techniques with each other and take a 'best of three' approach, and others compare indirect techniques against enzyme activity levels. An averaging of these surveys can be seen on the table below (Newcomer et al 1975, Howell et al 1981, Arola et al 1988, Peuhkuri et al 2000, Kurt et al 2003).

Type of test	False Positives	False Negatives
Breath Hydrogen	5/120	9/132
Blood Glucose	5/73	10/116

1.2 A table to show the averaged error rates derived from a series of studies reporting the accuracy of phenotype tests

1.2.2.3 Clinical symptoms of primary hypolactasia

If lactose is not digested by lactase, it reaches the colon without being hydrolysed, where it has an osmotic effect and causes diarrhoea. Also, any undigested lactose that reaches the colon is fermented there by colonic bacteria, and hydrogen is produced, which is absorbed in the blood stream and expelled in the breath. This forms the basis of the hydrogen breath test. The production of hydrogen and other gases may lead to flatulence, bloating and distention of the gut. Anaerobic bacteria convert the lactose to lactic acid and acetic acid, and other organic acids. The net effect is loose acidic stools. The set of symptoms

summarised above, which may themselves vary in type and intensity, are commonly known as lactose intolerance (for review, see Hollox and Swallow 2002).

Confusingly, although lactose intolerance often occurs when lactose is not absorbed, there is not a complete correlation between lactase non-persistence, lactose malabsorption and lactose intolerance (Jusilla et al 1969). A large proportion of non-digesters are able to drink nutritionally benificial amounts without exceeding a threshold at which symptoms are triggered, for example, 250ml of milk (Hussein et al 1978). There is some evidence that there is variation in this threshold (Bedine and Bayless 1973). Also, the fact that self-diagnosed 'lactose intolerant' individuals are often in fact lactase persistent may indicate a psychosomatic element to the symptoms (for example, Peukhuri 2000).

1.2.2.4 Disease associations of the two phenotypes

Aside from the symptoms of lactose intolerance, it is possible that there are other, less direct repercussions of both phenotypes. Several studies have investigated the clinical effects and disease associations of lactase persistence status, which may impact on relative fitness, with a possible selection pressure. Lactase non-persistence has been shown to have significant association with osteoporosis incidence (Birge et al 1967). It may be the case that, in Western populations, a higher milk intake provides some protection against the decalcification of the skeleton, although this observation seems unlikely to hold true across all human populations given the relatively low frequency of osteoporosis in populations with a high frequency of lactase non-persistence. Osteoporosis is unlikely to incur a significant fitness cost against the non-persistent phenotype given the comparatively late age of onset.

Similarly, although evidence exists for an association between high fresh milk consumption (enabled by lactase persistence) and coronary heart disease (Segall 1980), premature senile cataract (Simoons 1982) and hyperlipidemia (Sahi et al

1977), none of these are likely to have a significant impact on an individual's reproductive success.

These studies have generally not been controlled for ancestry / ethnicity, and have not studied the associations in a wide series of different populations worldwide. In terms of selective fitness, there is no evidence that continued expression of the enzyme has any adverse effects, except the inability to consume significant quantities of fresh milk in the diet, which may, indirectly, have a more general impact on health.

1.2.3 The distribution of the lactase persistent and non-persistent phenotypes

Appendix B3 provides a summary table of published literature for lactase persistence frequencies, and chapter 5 shows these data also plotted on a gradient smoothed contour map. A few general observations can be made regarding published frequencies of lactase persistence across the world. Within Europe, there is a Northwest to Southeast cline, with lactase persistence occurring at a significantly higher frequency in the Northwest. Lactase persistence occurs at intermediate frequencies in Central Asia and the Near East. In Africa, there is a general north to South cline, with higher frequencies of lactase persistence across the Mahgreb that diminish to low frequency or no evidence of the trait throughout most of sub-Saharan Africa. Isolated groups of nomadic pastoralists (such as the Beja) show comparatively high frequencies of lactase persistence, which disrupt this pattern. Nilo-Saharan groups such as the Nuer often display only a low to moderate frequency of lactase persistence despite a long history of pastoralism and milk drinking. Lactase persistence frequency is generally high amidst the Arabic populations of the Near East, and also Northern India but is almost absent in South and South-East Asia (for review, see Swallow 2003).

Several theories have been proposed to explain the modern distribution of phenotypes. One of the earliest observations was of an association between high frequency of lactase persistent individuals in a population group, and a long-

standing culture of milk drinking. This led to the development of a theory that suggested a relationship between the two, and the possibility that, in an environment where a milk drinking culture had developed, natural selection was responsible for driving the trait to high frequency.

In contrast, another theory suggests that a selective pressure favouring non-persistence in malarial regions of Africa is responsible for the high frequency of non-persistent individuals in the present day (Anderson and Vullo 1994). The authors suggest that infants who rejected breast milk due to an early onset of lactase non-persistence suffered mild riboflavin malnutrition; this might confer some resistance to malaria amongst African populations by providing an unfavourable environment for populations of malarial parasites. This theory is problematic because, as discussed, it is widely accepted that lactase non-persistence is the ancestral state in humans, making backwards selection less likely. Also, the comparatively late age of downregulation onset makes this unlikely, as does the fact that the selection pressure is limited to the weaning process when alternative food sources to breast milk are not readily available.

1.2.3.1 The Cultural Historical Hypothesis

The 'cultural historical hypothesis' was independently developed by Simoons (1970, 1978) and McCrackern (1971), who observed a relationship between frequency of lactase persistence and historical dependence on milk. They argued that, given an estimated time span of 6000 years for dairying and fresh milk drinking to be established in Northern Europe, there was sufficient time for adaptation to milk consumption to occur. In effect, this argument suggests that the cultural practice of milking co-evolved with the rise in frequency of lactase persistence. One difficulty with this hypothesis is the starting assumption that milking and using milk-based foodstuffs *per se* incurs a selective advantage on lactase persistent individuals. If a population were drinking fermented milk, or using milk based products that had been processed in someway, the lactose content is significantly reduced (for example, Alm 1982) such that the expression

(or absence) of lactase at high level made little difference to digestion of the milk and calorific benefit. It is also the case that not all lactase persistent individuals (or populations with a high frequency of lactase persistence) have high levels of milk consumption.

A later study by Aoki (1986) designed a stochastic model using the two genetically controlled phenotypes, (lactase persistence and lactase non-persistence), and two culturally determined phenotypes, (milk drinking and non-milk drinking), to investigate the relationship between selection, effective population size and the association between milk drinking and lactase persistence phenotype. His conclusion suggested that a complete correlation between milking and persistence should not be expected, and perhaps as significantly, that the selection coefficient to bring the frequency of the lactase persistence allele to that observed today in Northern Europe would need to exceed 5% given an effective population size of 100.

One of the difficulties with constructing a scenario of historic selection in Europe is that fresh milk drinking, even given an early date in the Neolithic, has not been around long enough to have influenced gene frequencies so completely unless selection or drift was very strong. Cavalli-Sforza and Feldman suggest that strong cultural transmission through the generations would be vital if an allele coding for lactase persistence was to reach high frequency in less than 300 generations (Cavalli-Sforza and Feldman 1989).

A recent study investigated the correlation between early Neolithic cattle farming sites and lactase persistence frequencies with variation in the genes encoding milk proteins in cattle (Beja-Pereira et al 2003). Seventy European cattle breeds were studied, with nonsynonymous mutations investigated in six key milk proteins to characterise the diversity of cattle milk genes. When the genetic variation of these loci from different cattle breeds was mapped against the geographical distribution of lactase persistence, and also Neolithic cattle farming

sites, a strong correlation was found, which, the authors suggest, provides good evidence for gene-culture evolution.

1.2.3.2 The Aridity hypothesis

Cook and Al-Torki (1975) supported the hypothesis that natural selection was responsible for the high frequency of lactase persistent individuals in some populations. However, they suggest an alternative selection pressure specific to isolated groups of Near Eastern and Middle Eastern nomadic pastoralists. They propose that, historically, nomads in arid regions might have been strongly dependent on milk as an alternative to fresh water. This hypothesis makes intuitive sense given the difficulties of finding and transporting fresh and uncontaminated water throughout desert regions.

1.2.3.3 The calcium absorption hypothesis

Flatz and Rotthauwe (1973) similarly suggested the possibility of more than one historic selective process occurring to explain the varying frequencies of lactase persistence observed today. They proposed the 'calcium absorption hypothesis', suggesting that the prolonged cloud cover in Europe, and its resulting low solar irradiation left some early settlers unable to synthesise vitamin D from sunlight, and therefore they were vulnerable to rickets and osteomalacia. Rickets in particular might have the effect of distorting pelvic development in women, with an associated higher risk of mortality in childbirth. Drinking milk might confer an advantage in this context since milk is an excellent source of calcium, and also lactose itself is claimed to stimulate calcium absorption (Flatz 1977). The authors suggest that this selection pressure explains the high frequency of lactase persistence in Northern Europe, while aridity is more significant in African and Near Eastern pastoralists, and that intermediate frequencies can be explained by historic (and modern) migration patterns. However, Suarez et al (1998) showed that lactase non-persistent individuals could, without complication, digest a dairy rich diet made from fermented food products, providing up to 1500mg of calcium a day.

1.2.3.4 Investigating these three hypotheses

Holden and Mace (1997) investigated which of the above hypotheses best explained the distribution of lactase persistence phenotypes, taking into account the confounding effect of the shared ancestry of many extant population groups. They identified a series of populations for which lactase persistence frequency data was available, also extracting information about pastoralism, milking and previous data describing genetic variation. Three different phylo-genetic trees were created; the first used F_{st} values taken from Cavalli-Sforza's study based on gene frequencies for a series of population groups (1994). The second tree modified this approach slightly by using Nei's genetic distance measure for the same data, and the third was constructed using language trees based on the classification of Ruhlen (1991).

Multiple regression analysis suggested that solar radiation and aridity could not explain the high frequency of lactase persistence in certain populations, but a history of pastoralism could. The authors concluded that a pastoralist lifestyle might have little deleterious effect for those with lactase non-persistence since processing milk foods reduces lactase. However, there would be opportunity for lactase persistent individuals to gain a selective advantage. Intermediate frequencies of lactase persistence were explained through drift and migration.

The study of Holden and Mace (1997) reviewed and statistically analysed the key hypotheses proposed, and also synthesised anthropological and genetic data, but it is possible that a different experimental design might provide a different answer. In considering 'solar radiation', for instance, the effect of changing cloud cover (which itself would vary considerably through time) was not considered, though this would, have some significant impact. A further difficulty was that of ensuring that there were sufficient genetic and anthropological data for all the populations studied; groups that did not appear in Murdoch's ethnographic atlas (1967) were not included in the analysis, since no ethnographic data was

available. Populations were also excluded if they were not included in Ruhlen's language atlas. Many Eurasian populations were not considered due to scarcity of raw data on lactase persistence status. It is therefore possible that further investigation with a more comprehensive data set would give a different answer.

Lactase has been frequently cited as a likely candidate for selection; however, the wide distribution of both phenotypes indicates that drift and migration have also played a major role in establishing modern day frequencies of the trait. The most probable scenarios, as discussed above, propose selection in the context of fresh milk drinking, combined with the effects of migration and drift. It seems necessary therefore to consider the background circumstances of the Neolithic and general history of pastoralist groups in order to better understand the influences that might have affected historical demographic events.

PART 3
The Agricultural Revolution, Pastoralism and Dietary Adaptation

Much of the archaeological theory in this section is greatly condensed, and represents only a small fraction of the ongoing work, debate and controversy in current archaeological research.

1.3.1 The Agricultural revolution

As early as 7500BCE some human populations began to domesticate a series of plant and animal species, ultimately developing an agrarian industry that was to spread rapidly throughout the populated world. Recently, the emergence of agriculture has been used to define the Neolithic (formerly defined by the presence of polished micro-liths), and so 'the Neolithic' often refers to a package of cultural developments rather than a chronological era. The causes of the shift are unclear, but several theories have been proposed: agriculture developing as a response to climatic change has been suggested, as has the notion that the inclination to manipulate and control the local environment comes from an innate human desire. Another theory suggests that a prehistoric population expansion necessitated more efficient food production, but conversely, it has also been suggested that agriculture enabled population expansion (for summary, see Price 2000).

Excavated sites associated with the transition from hunting and gathering to farming provide evidence for the development of artwork, agricultural techniques such as irrigation, weaponry, settlements and social stratification (for example, Renfrew and Bahn 1991). Archaeologists were surprised by the nature and scale of change, and so the transition itself has become known as the 'Neolithic' or 'Agricultural' Revolution (Childe 1936). Since there are such marked, observable changes reflected in the archaeological record, is it possible that this 'revolution' in human culture selected for changes in human biology as well?

1.3.1.1 Repercussions of the agricultural lifestyle

The cost of the cultural achievements of the Neolithic may well have been a series of new physiological stresses on the human body to which early populations were maladapted: new zoonoses such as measles developed from the cattle disease *rinderpest*, irrigation created environments for parasites such as schistosomiasis to thrive in, and ethnographies of modern hunting and gathering communities suggest that the Neolithic farmers and pastoralists may have increased their working hours, thereby placing an additional energy cost on the body (for example, Shostak 1983).

Most significantly, evidence of malnutrition in skeletal material from early agricultural settlements suggests that the new Neolithic diet was providing insufficient nutrition (Ulijazak 1993), which in turn raises the possibility that early human farmers were initially poorly adapted to their new diet. Reconstructions of the pre-farming diet suggest that a variety of foodstuffs were consumed: for example, sites in Europe reveal that large herbivores such as reindeer, mammoth, bison and horse, were hunted (Klein 1989). A variety of nuts and seeds were also consumed (Dumayne-Preaty 2001), and various grasses and grains (Harlan 1989). Conversely, most agricultural societies rely on a 'staple' carbohydrate, (such as rice in Asia, wheat in Europe, maize in the New World) and a more limited array of foodstuffs (Larsen 2000). This specialization in diet may have selected for an analogous specialization in digestion of the staple food, such that some communities became adapted to the new diet through selection pressures.

The nature of dietary change (and so the nature of potential selection pressures) is likely to have varied widely, being dependent upon both the local environment and the specific innovations of the early Neolithic communities. In some cases, there is evidence to suggest that a prolonged period of experimentation in domesticating plants and animals occurred alongside an economy based on

hunting and gathering, making the transition in diet far more gradual. For example, early pastoralist sites in Africa contain remains of wild grains, suggesting that the domestication of animals coincided with gathering wild foods (Barich 1992). This slower transition may have moderated the more extreme pressures resulting from the new diet.

1.3.2 The emergence of pastoralism

As discussed in 1.2.3, lactase persistence seems to be associated with a history of milk drinking. Milk drinking, in turn, is associated with animal husbandry as a distinct practice within agriculture. Consequently, the emergence of pastoralism is of relevance in determining the distribution of Neolithic milk drinkers who may have been subjected to the selection pressures that brought the lactase persistence trait to the high frequency found in modern populations, or whose migration may have been responsible for some of the intermediate frequencies observed.

1.3.2.1 The emergence of pastoralism in Eurasia

The first cultivated foods were observed in a series of sites in the Near East. This collection of early sites formed an arc of land known as the 'Fertile Crescent', starting in Palestine, progressing through the Levant up to eastern Anatolia then down to the valley of the Tigris river in Mesopotamia and bordering the region of West Iranian Zagros mountains. These farming communities appear to have practised a mixture of crop cultivation and animal husbandry, (for example, Scarre et al 1997) specifically utilising the following early domesticates:

Domesticate:	Site associated with domesticate findings:
Emmer wheat	Jericho, Aswad
Einkorn and Barley	Abu Hureyra
Sheep	Cafer
Pigs	Halian Gemi
Goats	The Zagros Mountains
Cattle	Anatolian Peninsula

Table 1.3 Summary of domesticates and examples of sites they have been discovered at (Scarre et al 1997)

Of the milking animals, the earliest dates for domestication are for goats, circa 7000BCE in the Levant (Legge 1996), and for sheep from the North East of the Crescent, circa 6600BCE. Cattle are recorded in Anatolia as early as 8000BCE (Grigson 1989), and in North Syria between 6000 and 5000BCE. In the Southern Levant, there is no definite evidence for cattle until 5000BCE (for example, at Tel Dan), after which they become plentiful in the record (Grigson 1989). Sheep and goats arrive at Asraq, East Jordan by 6000BCE, and it is likely that by this time trade, nomadism or extended contact between early communities were introducing some domestic species further along the fertile crescent from the Levant.

The early settlements of the Fertile Crescent were sedentary, but some mobility or trade between them and similar groups in Anatolia (modern Turkey) can be evidenced through the distribution of obsidian artefacts. Each volcanic source of obsidian has its own unique chemical signature, and so movement of specific pieces can be traced from their source. Various pottery styles can be used to chart the movement of specific cultures through regions and also, using radiocarbon dating, through time (Scarre et al 1997).

One of the most enduring theories regarding the spread of agriculture and pastoralism is the proposal that farming 'diffused' from the fertile crescent area to Upper Anatolia and across the Levant to Egypt, and from these two points, to the rest of Eurasia and Africa (Childe 1936). The degree to which the ideas of the Neolithic were exported from a central region as opposed to developing independently in different areas is a subject of some debate in archaeology, though it is generally agreed that the fertile crescent was one of the earliest centres of domestication and certainly the most influential for Eurasia.

1.3.2.2 The spread of the Neolithic to central and southern Asia

The final significant Eurasian centre of agriculture is in the Indus valley, where the flood plains provided fertile land for the early farming settlements. One of the earliest sites near the Indus River, Mehrgarh, has been dated at circa 6000BCE, and reveals evidence of Asiatic wheat and domesticated goats (Scarre et al 1997). It is likely that these were imported from communities near the Zagros Mountains in modern day Iran, which had domesticated these same species earlier, and there is strong evidence that the early societies of the area were influenced by the Near East (Tharpur 1973). By 5000BCE, discoveries of copper artefacts, turquoise from Iran and shells from the Arabian Peninsula suggest trade links running through Mehrgarh, a site slightly north of the Indus River. The Indo-Iranian border, and also Turkmenia, developed a secondary centre of civilisation derived from Mesopotamia; by the fourth millennium BCE, dozens of similar villages and towns were well established, and irrigation systems were used to water the land in the dry periods when the Indus river did not flood the plains (Biscione 1973).

Afghanistan was a crucial area linking early Mesopotamia with trade further east, and key sites such as Mundigak show evidence of prehistoric pottery with both Indus valley and Iranian style markings (Fisher 1973). Seistan, on the border of Iran and Afghanistan, was densely populated and a key trade route throughout the 4th Millennium BCE. It is likely that agricultural practices and possible nomadic caravans maintained strong cultural and economic exchange from Turkmenistan (for example, at Namazga III), through Seistan (shar-I-sokhtal), South Eastern Afghanistan (Mundigak III) through to the sphere of the early Indus cultures of the middle and upper Indus basin (Dales 1973).

It is believed that agriculture evolved independently in China and spread throughout the Far East and Australasia, but, since lactase persistence and milk drinking exist only at low levels in these areas, this thesis does not focus on that region.

1.3.2.3 The emergence of pastoralism in Europe

The earliest European agricultural settlements, contemporaries of the Near Eastern Ubaid culture, emerged in the Aegean islands and Greece by the beginning of the 7th Millennium BCE (Demoule and Perles 1993). Since there is no evidence for a prior human presence on the islands (Price 2000), this is likely to have been through colonization by Anatolian farmers. Traditionally, the arrival of agriculture throughout the whole of Europe has been viewed by archaeologists in this way, as effective colonisation by Neolithic communities. The indigenous hunter-gatherer groups were considered to be *'sparsely present, residentially mobile, socially amorphous and eventually overwhelmed.'* (Price 2000). This theory was based largely on the observation that most early farming communities were uniform in culture, suggesting the expansion of one cohesive founding group. Another source of evidence for the colonization of Europe has rested on language, and Colin Renfrew (1987) proposed that part of the Neolithic cultural 'package' was a proto-Indo-European language, a forerunner of the languages seen throughout Europe today. However, it is difficult to prove conclusively that there is an association between the spread of early Neolithic communities and the spread of language, and indigenous adoption of a language cannot be ruled out (Mallroy 1989).

Two main routes for colonization in Europe have been proposed: one from the Anatolian plains across the Aegean to the Balkans, and then north and west; the other along the northern shore of the Mediterranean and eventually upwards into Central Europe (see fig 1.3).

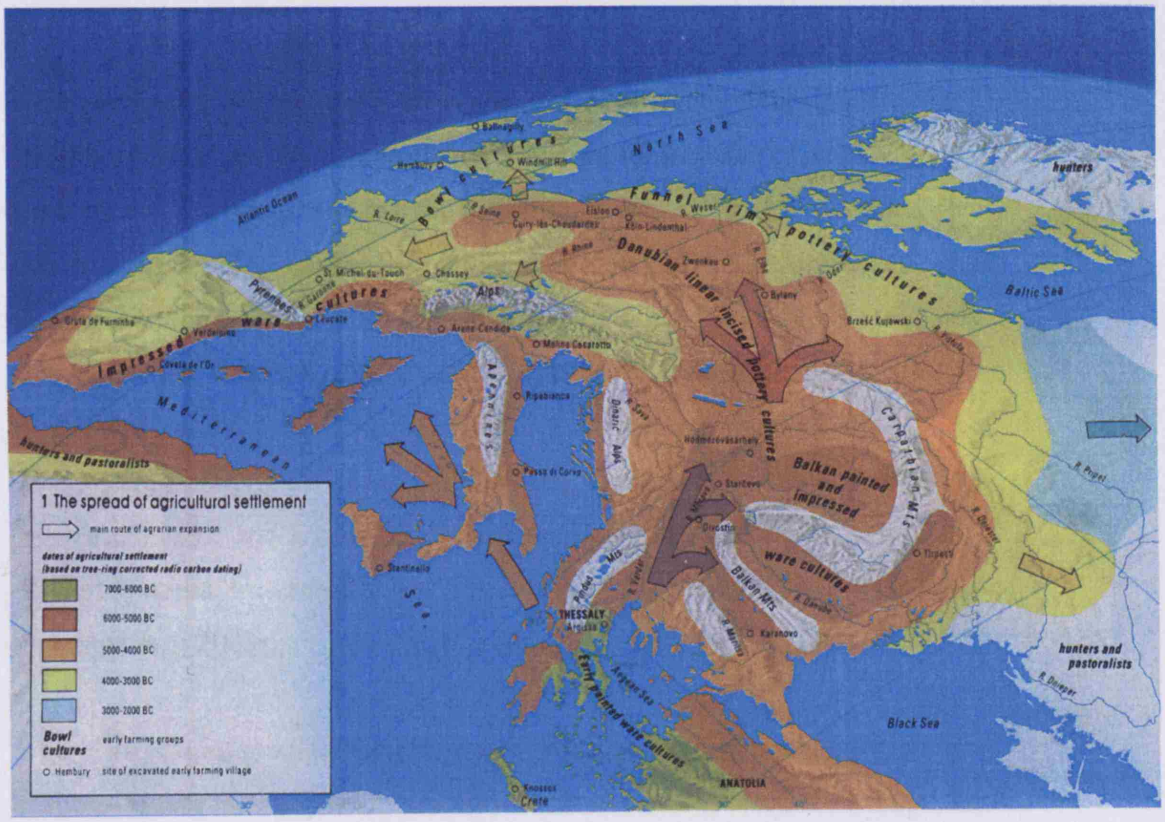


Fig 1.3 - A map to show the expansion of Neolithic culture across Eurasia.
This map, taken from the Concise Times Historical Atlas (1994) shows the conventional pattern of expansion of Neolithic peoples from Anatolia

This pattern of transmission has been referred to by Ammerman and Cavalli-Sforza as a 'Wave of Advance' (1973), due to the steady progression of Neolithic culture, which they suggest, based on radiocarbon dates, occurred at a rate of 1km per year. Radiocarbon dating provides evidence that 750,000km^2, almost all of central and southern Europe, had been settled by farming communities by 5000BCE (Bogucki 1988). Interestingly, after an initially successful diaspora, the spread of farming appears to have simply stopped, and a 'stop line' has been mapped which runs from the Ukrainian steppes, across Northern Poland and Germany, into eastern France, and then back towards the upper Rhine. Neolithic ground stone tools found in foraging sites in Southern Scandinavia suggests that this line was traversed by traders (Fischer 1982), and possibly also by nomadic pastoralists searching for new grazing ground (Bogucki 1988).

Most relevant to this project is the correlation between this stop-line and the region in central Europe where the areas of high frequencies of lactase persistence appear to drop from the higher frequency in Northern Europe, for example, from 96% in Ireland (Fielding et al 1993), to the lower frequencies of Southern Europe, for example 50% in Greece (Kanaghinis et al 1974). This is opposite to the direction of migration suggested by traditional archaeology, in which pastoralism was imported from the Near East in a wave of advance. Above the stop line in Northern Europe, the farming villages of Britain and Scandinavia were established only by 3000BCE, yet it is in modern Scandinavia and the British Isles that lactase persistence occurs at near-fixation point (for review, see Swallow et al 2003).

There are several possible explanations for this observation: that the mutation responsible for lactase persistence may have occurred in Northern Europe, and subsequent back migration carried it across Europe to the South and beyond; the trait may have evolved elsewhere, but selection and/or drift acted most strongly in Northern Europe to bring lactase persistence to high frequency; there may be more than one mutation event bringing lactase persistence into high frequency in Northern Europe and the rest of the world.

Recently, archaeological finds have suggested that the colonization model may be over-simplistic. Evidence of domesticated animals has been found in Scandinavia (Jennbert 1984), the Mediterranean (Geddes et al 1989) and in Ireland, suggesting that Mesolithic communities may have experimented in domestication. The most extreme advocate of this idea, Dragoslav Srejovic, uses discoveries from Lepanski Vir and Vlasac in the Balkans to suggest that experiments in animal (specifically cattle, dog and pig) and plant domestication were developing contemporaneously with the Near East. Although this scenario is not widely accepted, there is a sufficient body of evidence to suggest that Mesolithic foragers intensively manipulated their local resources (for example, Rowley-Conwy et al 1987). It is possible therefore that some form of proto-

pastoralism was present prior to the arrival of Neolithic farmers from the Near East. This would make the first scenario, that is, the emergence and spread of lactase persistence in Mesolithic Northern Europe, possible, though still unlikely.

Excavated sites in Northern Europe show a long-term continuity in lithic and ceramic styles, which may suggest the indigenous adoption of Neolithic culture and ideas by local groups (possibly influenced by contact with central European farmers) rather than total population replacement (Price 2000). Osteoarchaeological studies suggest few differences in morphometrics between Mesolithic and Neolithic skeletal remains (Schwidetsky 1973), and a study of craniometric features observed in two Mesolithic groups and in one Neolithic group in central Portugal showed no significant differences, despite a time difference of 3000 years and differing ecological climate (Jackes et al 1997).

Given this new archaeological evidence, it seems likely that the spread of farming in Europe is a complex and variable combination of both demic diffusion, where immigrants colonize the land, and also of indigenous adoption of the ideas of the Neolithic.

1.3.3 The emergence of Pastoralism in Africa

In Africa, good evidence exists both for an independent domestication for animal husbandry, specifically cattle farming, and an imported one. If pastoralism spread to Africa from the Near East, or even the other way around, do modern pastoralists in Africa share a comparatively more recent genealogical history with pastoralists from Eurasia compared to their neighbours?

1.3.3.1 Evidence for independent domestication of cattle and independent development of pastoralism

The earliest date proposed for domestic cattle in Africa (at Bir Kiseiba in Egypt) is 7500BCE, which predates Near Eastern cattle domestication and therefore indicates that the domestication of cows took place independently there (Gautier

1984). Nabta Playa has also yielded cattle remains which may be domestic, 6840 +/-90 BCE (Gautier 1980, Weondorf and Schild 1980). The evidence from both studies depends partly on the interpretation of osteological material, which the authors suggest show changes in size, armoury and build indicative of domestication. The other component of the argument is based on the observation that the low rainfall in the Nabta Playa area created an arid environment, and that cattle would be unable to survive without human aid (Neumann 1989). This theory is supported by the complete absence of comparable taxa in the archaeological record (Close and Wendorf 1992). Dating of sites is crucial to the argument, since by 5700 – 4500BCE the same site yields uncontroversial cow domesticate remains (Wendorf et al 1996). By 5000BCE, ovicaphrines are also found in North East Africa (Vermeersch et al 1996) but, by this time, contact with west Asiatic herders is likely.

A more robust piece of evidence for independent domestication comes from phylogenetic analysis of modern cattle mtDNA, which suggests a genetic separation of 275-117 thousand years ago for Indian (*Bos Indicus*) cows, and Euro-African cows (*Bos taurus*), and, interestingly, that the most recent common ancestor for African and European cows is 26-22 thousand years ago (Bradley et al 1996). This date is far earlier than any known estimate of domestication either in the Near East or Africa, and so suggests that the wild Bos species had diversified, and were subsequently domesticated independently in the two continents. However, if female African Bos were crossbred with a Eurasian species, further study of the Y Chromosome or autosomal DNA analysis might reveal a different diversification date. A small founding Euro-Asiatic herd might also have been augmented and eventually replaced by wild breeds. This possibility, which Bradley et al (1996) discuss, is credible given the different environmental stresses on the African continent that indigenous species might have been better adapted to cope with.

1.3.3.2 The spread of pastoralism in Africa

Some of the earliest pastoralist sites have been described as 'Steinplatze', or 'Stone place' sites, as they consist of a concentration of stones, hearth rings and waisted tethering stones. No substantial material culture has been found from these places, and it has been suggested that they are ephemeral stopping points used by early nomadic pastoralists (Gabriel 1973, 1978). They are found in Eastern Egypt and the Sudan, Libya, Northern Chad and Algeria, along with animal remains. Rock art depictions of these stones in Libya suggests that they were used for tethering cattle, (Castiglioni and Negro 1986), though this is difficult to prove conclusively.

The spread of animal domestication can be observed by dating (where possible) pastoralist sites, the earliest of which are in the Mahgreb and in Northern Chad, such as Enneri Bardague, where cattle, sheep and goat remains have been dated at 5455BCE (Gautier 1987). In Wadi Ti-n-Torha in Northern Libya, a median date for cattle has been established at 4000BCE with the earliest date at 5070+/-60BCE (Barich 1987), and at Ti-n-Hanakaten in Algeria, a median date of 4650 +/-90BCE (Aumassip and Delibrias 1982-3) for cattle. By 4000BCE, cattle were also well established in pre-dynastic Egypt. Between 4500 – 2500BCE, pastoralism had spread to the Hoggar Mountains, and to Kadero, North of Khartoum by 4000BCE. At approximately 3000BCE extensive migration as a result of the increasing desiccation of the Sahara and the tsetse fly barrier spread cattle, and probably sheep and goats as well, deeper into Africa; there is evidence of migration across the Sahel corridor in 3760BCE to Adrar Bous in Niger, and the West African River basins start to show sites associated with pastoral groups by 2500BCE. Gaji in Kenya has sites c2000BCE, and Zambia, the South and the Cape Province around 400CE (for review, see MacDonald 2000). Although cattle are predominant in the archaeological record, it has been suggested that goats and sheep were also transported, and some evidence, such as naturally mummified sheep in Lower Nubia (Kerma site) have been dated between 2500 – 1700BCE (Chaix and Grant 1987).

The development and spread of pastoralism within Africa is very different from that of Eurasia: Sahelo-Saharan studies showing evidence of livestock without contemporaneous evidence of crops or field clearing suggest that in most of Africa, nomadic pastoralism predates cereal agriculture, in some places by as much as 2-3 thousand years (Macdonald 2000). The implications of this are significant, as they undermine the central dogma of conventional archaeology, which asserts that sedentism is a precondition for domestication (for example, Flannery 1969, Reed 1977). The mobility of pastoralists would increase the speed of their contact with other groups and might increase the rate at which animal husbandry and the broader concept of domestication was spread throughout Africa.

1.3.4 Milk drinking

The advantages of milk drinking (assuming some degree of lactase persistence already exists in a population) are clear: a hygienic, portable and continuously available alternative to fresh water, which, as discussed, might have historically been important for nomadic pastoralists in arid regions where water sources could not be guaranteed (Cook and Al-Torki 1975). Milk itself is highly nutritious; it has three main constituents, water, fat (mainly triglycerides) and non-fatty solids, including casein, albumin, proteins and lactose (Egan et al 1981). Table 1.4 summarises the key components of milk in a series of different mammals (Egan et al 1981). Human milk has a comparatively high lactose level compared to most other mammals, perhaps because of the extent of neotany of human babies.

Animal	% Water	% Fat	% Lactose
Human	87.6	3.8	7.0
Cow	87.2	3.6	4.9
Ass	89.8	1.4	6.2
Buffalo	82.4	4.7	4.6
Ewe	80.6	5.4	4.7
Goat	87.8	3.5	4.1
Llama	86.5	3.9	5.6
Mare	89.8	2.0	6.1
Reindeer	66.1	10.1	2.5

Table 1.4 Water, fat and lactose concentrations in milks of a variety of species (from Egan et al 1981)

The difficulty with investigating ancient milk-drinking practice is that it is almost impossible to demonstrate conclusively from the archaeological record. Attempts at identifying milk-use through trace residues of milk proteins are still in their early stages, though a study by Craig and Collins (2000) reported a new method, which dissolves the ceramic to liberate the milk proteins, and then uses antibodies to detect them. This technique was used to infer an early Iron Age dairying economy at Cladh Hallan, South Uist in the Outer Hebrides (Craig et al 2000). Where milk residues are observed, it is difficult to tell whether the milk used was fresh, or whether it was treated in some way, which might have reduced the lactose content.

Milking images, such as rock art, are sometimes used as evidence to suggest that milking animals has occurred in an ancient population. However, dating such images and associating them with specific archaic societies is problematic.

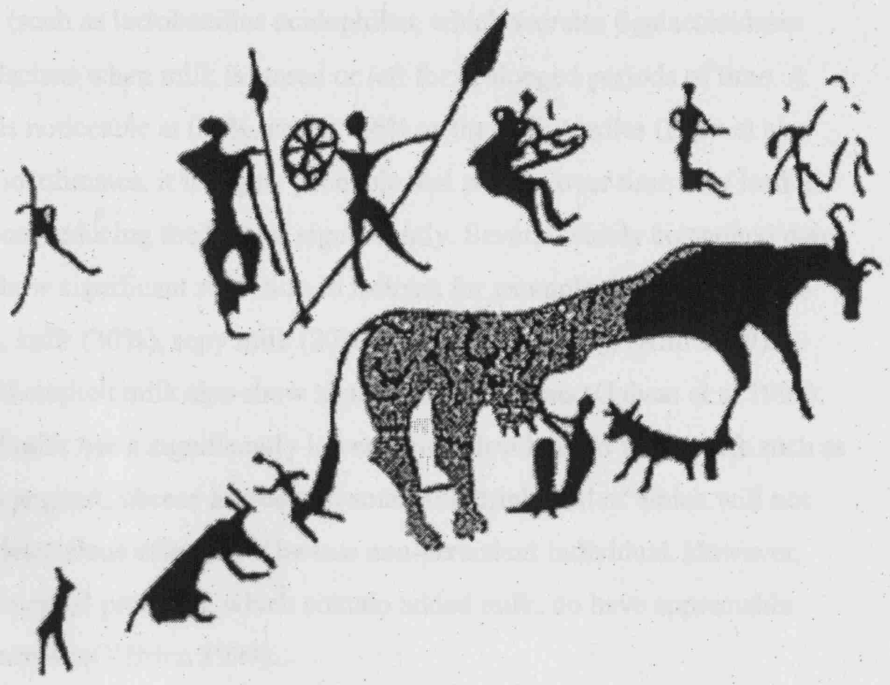


Fig 1.4 An example image of a rock painting found in a cave in Eritrea, illustrating early milking practice, dated 3000 BCE approximately.
Blench (2000), after Graziosi (1964a)

Modern day fresh milk drinkers are not always easy to identify from anthropological literature. Murdock's ethnographic atlas (1967) suggests that, of 28 African pastoralist groups with a subsistence dependency of 46 – 100% on animal husbandry, all practised milking. Most of these groups herded cattle, although some in North Africa milked camels. In Eurasia, Murdock's data is less complete, and diversification and specialisation in the agriculture industry makes identifying milking practice more difficult to identify conclusively.

Part of the problem is that pastoralists who routinely milk their animals may not consume fresh milk, but instead may produce a variety of different food stuffs from processed milk, all of which are likely to have reduced lactose levels (Suarez et al 1998). The natural process of fermentation, which occurs if milk is stored for a period of time, especially at a warmer temperature, may well have been practised in prehistoric times. Milk contains 0.14% lactic acid, which increases in concentration when fermentated, due to the activity of micro-

organisms (such as lactobacillus acidophilus, which secretes βgalactosidases including lactase when milk is stored or left for prolonged periods of time. A sour taste is noticeable at 0.3%, and at 0.6% as the milk curdles (Egan et al 1981). In hot climates, it is highly probable that storage over time may lead to fermentation, reducing the lactose significantly. Several widely consumed dairy products show significant reduction in lactose: for example, buttermilk (26% reduction), kefir (30%), ropy milk (20%) and yoghurt (50%), (Alm 1989). Cheese and curdled milk also show significant reductions (Galvao et al 1995). Fermented milk has a significantly lower lactose level, as do food stuffs such as traditional yoghurt, cheese and certain sour milk drinks, all of which will not have any deleterious effect for a lactase non-persistent individual. However, some commercial products, which contain added milk, do have appreciable levels of lactose (O'Brien 1999).

Even those cultures that currently do use fresh milk in their diet may not have a high frequency of lactase persistence; for example, among the Nuer where lactase persistence occurs at a frequency of 17% (Bayoumi et al 1982) despite a long history of milking and fresh milk drinking. For some pastoralist populations with low or intermediate frequencies of lactase persistence, the presence of the trait may have resulted from admixture with groups where lactase persistence frequency is high rather than a historic selection pressure.

PART 4

Population Genetics: questions and methods

Population genetics provides a tool for inferring prehistoric demographic events that may explain the current distribution of lactase persistent individuals. The study of polymorphic loci within and between sample groups can be used to describe and investigate evolutionary processes. These processes include genetic drift (in which alleles are fixed or lost in a population through stochastic processes), gene flow (for example, due to migration), changes in effective population size (bottle necks and expansions, which directly effect the rate of drift) and natural selection.

1.4.1 Investigating human variation

1.4.1.1 Variation between modern human population groups

Despite a wide variation in habitats, lifestyle and culture, human beings are a fairly homogenous species, with comparatively little divergence between population groups. Rosenberg et al (2002) used 377 autosomal microsatellite loci to categorise 1056 individuals from 52 populations, and showed that diversity within groups accounted for 93-95% of total variation in humans. Even so, it was still possible to group the human populations studied into six genetic clusters, five of which correspond to continental divisions: sub-Saharan African, Eurasia, Oceania, East Asia and America. This finding was concordant with that of an earlier study by Lewontin (1972) who used blood groups to show that the vast majority of variation existed within populations and identified a similar, broad clustering of groups.

Studies using non-recombining loci, namely mitochondrial DNA and the Y Chromosome have shown greater inter-population diversity, which can be used to distinguish discrete population groups. For example, 12 biallelic polymorphisms and 6 microsatellite markers on the Y Chromosome were used to resolve haplotypes in a series of North Welsh and central English groups (Weale

et al 2001). Using statistical analysis of these markers, it was possible to determine that the English towns were genetically very similar, and the two North Walian groups were significantly different both from each other and from the English. Interestingly, samples collected in Friesland were statistically indistinguishable from the English group, suggesting a comparatively recent common ancestor. The authors hypothesised that the data was best explained by a past mass migration event that brought a substantial number of Y-Chromosomes to central England; specifically, the Anglo-Saxon migration supported by historical evidence (Weale et al 2001).

1.4.1.2 Use of genetic markers to measure diversity

This project uses three main types of polymorphic markers: the first, single nucleotide polymorphisms (SNPs), are defined here as 'polymorphic' when the least common of the two alleles observed exists at a frequency of greater than 1% in the population groups under consideration (Cummings 2000). SNP diversity is likely to vary between different chromosomal regions, for example, due to molecular causes (such as the greater mutability of CpG rich regions). Point mutations are thought to occur on average at a rate of 2.3×10^{-8} per locus per generation (Nachman and Crowell 2000).

An insertion/deletion polymorphism (InDel-intron1) is also described in this thesis; insertion or deletions of sequence, usually between 1 and 20 base pairs are thought to occur uniquely in the molecular history of a DNA region, at a frequency of 2.3×10^{-9} (Nachman and Crowell 2000), though the frequency of larger InDels is likely to be higher.

The third type of marker, microsatellites, are a class of variable number tandem repeat polymorphism (VNTR), with a repeated DNA sequence motif between 2 and 5 base pairs long, (for example, GA_N). Microsatellites have a comparatively high mutation rate, estimated at 0.0012 per locus per generation (Weber and Wong 1993, Brinkmann et al 1998), though a heterogeneous mutation process is

likely (for example, Ellegren 2000). They can provide greater resolution for determining diversity both within and between populations. Compound SNP-microsatellite haplotypes can be used to maximise information when determining selection pressures and demographic modelling.

1.4.1.3 Linkage Disequilibrium (LD) and Haplotypic diversity

Linkage disequilibrium (LD) refers to the observed association between alleles in a population study, where two markers appear together on the same chromosome more frequently than if they were segregating at random (that is, linkage equilibrium). Markers that are located close together on a given chromosome are less likely to be separated during meiosis than loci that are far apart. However, many studies have shown that LD across a particular region cannot be explained as a function of distance alone (for example, Taillon-Hiller et al 2000). Instead, LD appears to extend across regions of a chromosome in a 'block like' structure, with long sequences of DNA showing high levels of linkage disequilibrium punctuated by 'recombination hotspot', or areas where LD breaks down (for example, Stumpf and Goldstein 2003). These blocks can be used to locate candidate genes or regions that may be associated with disease, or other genetically controlled traits where the causal mutation has not yet been identified (for review, see Petes 2001).

The factors affecting the structure of LD are complex; recombination itself may not be uniform, and there is some evidence to suggest that location on a chromosome affects the rate, with conserved regions of low LD nearer the centromere (Petes 2001, Philips et al 2003). Stumpf and Goldstein (2003) suggest that, on the basis of differing demographic history, some human populations will show a block-like structure of LD, whereas others will not, even if heterogeneity of recombination rate is taken into account. Observing LD also depends on the data set used: a lower density SNP map may not inform about variants with low minor-allelic frequency, such as might be found in older populations with a greater degree of recombination, or in an admixed populations

(Goldstein et al 2003). The fact that, in addition to molecular causes, demographic events affect observed LD may complicate disease mapping, but provides a useful tool in population genetics.

Some studies suggest that LD decays as the 'age' of a population increases: thus sub-Saharan African groups, which are thought to be the oldest human populations, show comparatively low LD (Lonjou et al 2003). One study investigated a series of Finnish groups with different histories, and those known to be recently settled in Finland demonstrated a greater degree of LD than populations thought to be amongst the first settlers in the area (Varilo et al 2003). LD is also affected by admixture (Chakraborty and Weiss 1988), and it may be possible, conversely, to infer the proportion of admixture using LD (Bertorelle and Excoffier 1998). There are also tests to identify selection signatures that utilise LD (for example, Slatkin and Bertorelle 2001), and these will be discussed in more depth in section 6.1.

In regions of high LD, very few haplotypes (defined here as a particular combination of alleles along a chromosome) are present in the population. Recombination increases haplotype diversity by juggling combinations of alleles, and boundaries of recombination (hot spots) break down the length of a cogent haplotype. Since this is also affected by population history and selection, Therefore, the frequency of a particular haplotype, and its length across a region of DNA in comparison with other haplotypes, may be informative about population demographics. Several methods exist for constructing haplotypes and also for measuring LD (see 2.6).

1.4.2 Identifying historic Demographic events

1.4.2.1 Natural selection

Individuals who are advantaged in some way by a given phenotype are more likely to pass their genes on to the next generation. An advantageous phenotype will exist in higher frequencies in the environment in which it is favourable.

Much of selection is context-specific; a phenotype that is beneficial in one environment may lose its advantage in another, and may even, given a change of context, be detrimental to an individual's survival.

Although natural selection acts on the phenotype of individuals, the particular trait that has incurred reproductive advantage must be under genetic control, or it cannot be inherited (Bamshad and Wooding 2003). This central dogma focuses evolutionary study not only on the phenotype of an individual, but also on the success, spread and frequency of a particular allele. Natural selection can impact in various ways: as an increased ability to survive to reproductive age; as an enhanced ability to attract a mate (or 'sexual selection') and as differential reproductive success, which may be due to increased fecundity or to a greater ability to keep offspring alive.

1.4.2.2 Different types of selection pressure

Selection can be described as 'positive', by which it is meant that a particular phenotype is favoured, or 'negative', as in the case of a deleterious phenotype that is comparatively unsuccessful or even harmful. At the DNA level, selection acts on sequence differences that arise from mutations.

'Diversifying selection', as its name suggests, increases diversity, as in cases where heterozygotes carry an advantage that homozygotes for either allele lack. The most famous example is humans is sickle cell anaemia (OMIM 603903). Heterozygotes are thought to be under positive selection in malarial regions in Africa, since they are less likely to be infected by the malarial parasite, yet do not carry the more serious (and sometimes fatal) consequences of the sickle cell trait in its homozygous form (Haldane 1949).

Balancing selection similarly maintains diversity in a population. In the case of human immune recognition systems, many alleles are possible at a number of loci, and frequencies of these fluctuate. Here, several factors are important, such

as the interaction between pathogen and host, and the evolving dynamic of the population group as a whole. If an allele reaches high frequency in a population, it may well, after some time, lose any advantage, since pathogens will themselves become selected for an ability to evade recognition by individuals carrying that allele. Here, heterozygosity across MHC loci and also in the population as a whole increases the number of pathogens that can be recognised by the immune system, increasing the chances of surviving infection (Hughes and Nei 1988).

1.4.2.3 Tests for natural selection based on neutrality

In 1968, Kimura challenged the conventional view that selection was the only driving force behind the diversity he observed, and, proposed the 'neutral theory'. He proposed that most mutations are selectively disadvantageous, and therefore lost, or are selectively neutral (Kimura 1968). Alleles that do affect function and cause amino acid changes need to perform as well as the ancestral form to be considered neutral. It is also the case that mutations incurring slightly deleterious changes can behave as neutral alleles (Ohta 1992). The frequency patterns of most polymorphisms can therefore be considered as 'neutral' in that they frequency reflects a dynamic balance between the rate of genetic drift and mutation rate.

Tests for selection based on the neutral theory use frequencies of neutral distribution. These can then be used to provide a 'null hypothesis' against which the frequency distribution of alleles that may be under selection can be compared. One of the simplest types of these, 'codon-based' selection tests, compares mutations that produce synonymous or non-synonymous changes at a translation level. The test can be summarised by the following equation, $w = d_n / d_s$, where d_n are non-synonymous and d_s are synonymous changes, expected under nutrality. If $w > 1$, positive selection has occurred, if $w = 1$, no selection, and if $w < 1$, purifying selection (for review see Yang 2001).

This test was used successfully by Rooney and Zhang (1999), who investigated d_n and d_s in a gene coding for a protoamine involved in binding sperm DNA during spermatogenesis (P1). Non-synonymous mutations were shown to be significantly higher in hominoids and old world monkeys, suggesting positive selection (Rooney and Zhang 1999). Many tests for selection based on neutral expectation are of more use in determining divergences between and within species, such McDonald and Kreitman's 'G' test, which compares genomic regions within and between species for non-neutral patterns of polymorphism and diversity (McDonald and Kreitman 1991). Further tests of this kind similarly rely on the pattern of diversity compared within and between species (McDonald 1996, 1998).

Looking specifically at humans, neutrally based selection tests can now use genomewide datasets, such as those from the Human Genome Project. This resource has grouped data from three populations, African-American, European-American and Asian, with comprehensive genotypes available for over 28,000 SNPs. This has led to an increase in the data available and has enabled a more extensive genomic comparison of neutrality (for example, Akey et al 2002, Kayser et al 2003). It is possible to compare the frequency distribution of a locus of interest with that of the null distribution generated by other loci from the same samples and deviation from this expected differences in allele frequency can be used to identify polymorphisms under selection (for example, Akey et al 2002). However, the origin of the three groups sampled is problematic, given both the likely degree of admixture and also the tremendous diversity in Asia and Africa. Given these considerations, the populations cannot be considered representative of the African, European and Asian continents as a whole. Another issue is of 'selective sweep', in which a neutral allele (that is, an allele with no affect on fitness) is in LD with one under selection, that allele demonstrates a non-neutral pattern of distribution.

1.4.2.4 Tests for natural selection based on frequency spectrum

Measuring the level of diversity in a population in and of itself can be informative about selection. Under conditions of no selection, the expected diversity (θ) can be summarised as follows: $\theta = 2nN_e\mu$, where n = heritable copies of the locus per individual and μ = mutation rate per site per generation. N_e signifies effective population size, defined as the stochastic process of generational sampling, which may result in the descendants of each generation being non-randomly distributed. This formula reflects the rate of loss of alleles by drift, as determined by N_e, and the rate of introduction of new alleles into the population, as determined by μ. N_e is sometimes used to refer to an 'ideal' population, with no interlapping generations, and an equal male: female ratio and constant size (Wright 1931), and so although N_e has a relationship to actual population size (N), it often contrasts it.

One of the best known tests of this type, Tajima's 'D', uses two estimates of diversity in a population set. One is based on the number of polymorphic sites in region of DNA sequence in a sample; the other is based on the number of pair wise differences in the data set. Under conditions of no selection, both estimates should give the same result, and D = 0. However, if there are significant differences, a negative D value suggests positive selection, and a positive D value suggests balancing selection. This test was used to investigate diversity in the FOXP2 gene, which codes for a transcription factor used in speech and language development (Zhang et al 2002). The authors showed that there was a lower observed diversity in the intron of this gene than in equivalent non-coding regions, with a D value of −1.36, indicative of natural selection (Zhang et al 2002).

1.4.2.3 Tests for natural selection based on intra-allelic diversity

Tests based on intra-allelic diversity compare the joint distribution of allelic variation and frequency between a series of haplotypes in a population, where one haplotype is thought to be under selection and the others are not (Slatkin and Bertorelle 2001). There are two key assumptions for the model: first that, in the absence of selection, new alleles are likely to exist at low frequencies in a population, since insufficient time has passed for them to increase in frequency through drift. The second assumption is that there will be less intra-allelic variation associated with new alleles that are not under selection, since less time has elapsed for it to accumulate. Therefore, an allele showing low variation can be considered 'young', and, under neutral conditions, is expected to exist at comparatively low frequency. Tests based on intra-allelic variation use a series of linked markers in a particular genomic region as a proxy for the age of the allele under investigation. The joint distribution can then be used to determine whether a particular allele departs from neutral expectation. If the observed variability of an allele is inconsistent with the observed frequency of that allele, that is, if a relatively new allele exists at a high frequency, it should be considered a potentially candidate for selection (Slatkin 2001).

Using this approach, two different measures of intra-allelic variability can be used. One is based on recombination, the other, on microsatellite diversity. As discussed, over time recombination increases haplotype diversity by generating new combinations of alleles and by reducing the length of haplotypes. If an allele rises to high frequency sufficiently fast, there is insufficient time for recombination to break down the haplotype. This can be observed by measuring the regional linkage disequilibrium for a haplotype or the number of recombinants at a linked binary marker. Microsatellites linked to a SNP allele or a particular locus of interest are likely, with their high mutation rate, to accumulate new mutations and thereby increase intra-allelic diversity over time.

In 1998, Stephens et al published allele frequency data for an InDel polymorphism in the CCR5 gene, with two linked microsatellite markers. The deletion in the CCR5 gene removes a chemokine receptor on lymphoid cells. Many pathogens, most notably the human immunodeficiency virus (HIV-1) uses a chemokine 'coat' in order to gain access to a cell. Without the receptor, the retrovirus is unable to gain entry and to replicate (Stephens et al 1998). Using closely linked microsatellite markers to measure intra-allelic diversity, Slatkin and Bertorelle (2001) showed that the allele is young for its frequency of 10% in Europeans. This suggested a departure from the neutral expectation, indicative of selection. It was hypothesised that a historic plague, most probably, the Black Death, was survived preferentially by carriers of this allele.

The intra-allelic approach was also applied successfully by Sabeti et al (2002), who used extended regions of disequilibrium as a measure of intra-allelic diversity. They investigated an allelic variant of glucose-6-phosphate-dehydrogenase, which was suspected of providing a selective advantage against malaria in certain populations. They were able to demonstrate that the haplotype containing the protective allele was extended over a longer region than haplotypes carrying the ancestral allele, suggesting it had risen to high frequency before recombination and further mutation could accumulate and disrupt the linkage disequilibrium of the extended haplotype. More recently, Bersaglieri and colleagues used a variation of this technique to identify a possible selection signature for lactase persistence, which will be discussed in depth in chapters 6 and 7.

1.4.2.4 Confounding factors

An allele can also reach high frequency quickly through drift. An ancestral founder effect, might, for example, appear similar to a selection effect when using intra-allelic variability tests. One potential solution to this problem is to have an '*a priori*' hypothesis (Kreitman 2000) where selection may reasonably be expected to have influenced a trait. Lactase persistence is a good candidate

here, since the observed correlation between pastoralism and a higher frequency of lactase persistent individuals is strongly indicative of natural selection (see Simoons 1970, 1978 and McCrackern 1971). Another approach would be to try a combination of different tests. For example, using an intra-allelic diversity model combined with a model looking at allele frequency differences between populations.

Lactase persistence has already been the subject of many studies looking at the relationship between a culturally determined trait, milking, and a genetic pattern. The new genre of selection tests available makes lactase a natural candidate to examine the cultural-genetic process of evolution.

Part 5: Aims

The general aim of this thesis is to explore the hypothesis that the change in human culture from hunting and gathering to pastoralism during the Neolithic is reflected in modern patterns of variation in the gene coding for the enzyme lactase phlorizin hydrolase.

The specific aims of the thesis are as follows:

To determine the haplotypic background of the recently described C–13.9kbT polymorphism and the G–22kbA polymorphism (Enattah et al 2001), and an Insertion / Deletion polymorphism occurring in intron 1 of the lactase gene.

To investigate their relationship to the core lactase haplotype markers

To investigate the extent of linkage disequilibrium in and around the lactase gene.

To establish the distribution of the -13.0kb*T, -22kb*A alleles and the A haplotype (Harvey et al 1998, Hollox et al 2001) in a series of population samples from Africa and Eurasia.

To investigate the association of these markers with pastoralism, milking practices and known frequencies of the lactase persistence phenotype in a series of populations from Africa and Eurasia.

To construct microsatellite haplotypes, using family samples, from a series of populations, and to investigate the intra-allelic variation between core lactase haplotypes in these populations.

To investigate, using tests based on intra-allelic variation, whether a historic selection event is responsible for the high levels of lactase persistence in European populations.

To investigate the relative contributions of selection, migration and drift in shaping present day distributions of lactase haplotypes.

Chapter Two
Methods

Methods

2.1 Samples

All 4024 DNA samples used in this PhD thesis were collected with informed consent, and provided by either The Centre for Genetic Anthropology (TCGA) or the Galton Laboratories (see acknowledgements for individual contributors). Ethical and legal approval was obtained as appropriate for each country where samples were collected[4].

2.1.1 TCGA samples

TCGA kindly made available 3261 Sub-Saharan African and Eurasian DNA samples collected from unrelated adult males, and anthropological data was available for each detailing: birthplace, self-declared cultural identity, and first and second languages for each individual sample donor, their parents and paternal grandfather, and maternal grandmother. A series of family samples comprised both parents and at least one child and were collected from Irish, English, German, Algerian, Armenian, Ashkenazi Jewish, Ethiopian (Amharic) and Kuwaiti populations. Background information on self-declared ethnicity of the donor was available for each sample.

2.1.2 Galton Laboratory samples

The Roma, North and South Indian, San and Bantu samples were collected by various donors working with the Galton laboratory (see acknowledgements and Hollox et al 2001), and the Northern French family population was collected by the Centre d'Études du Polymorphism Humain (CEPH). The Finnish samples were collected and phenotyped in Dr. Riitta Korpela's laboratory using the 'Gold

[4] The UCL Joint Committee on the Ethics of Human Research approval reference number for TCGA, where this project was based is: 99/0196. The UK and Finnish patient samples used have an approval reference number: 01/0236

Standard method' (Peukhuri 2000). Background information on ethnicity was provided, where available, on request from sample collectors.

2.1.3 Categorisation of samples

Ethnic categorisations are often used to identify distinct groups for analysis in population genetics. These classifications must be approached with some caution, since they describe a complex set of cultural associations which have different degrees of importance depending on where, when and from whom the samples are taken (for example, Chapman 1993). In the course of this thesis, sample donors described their ethnicity or cultural affiliations in various terms: by family or 'clan', home town, nationality, country or continent, and a basic priority was to ensure that the same level of resolution was used wherever groups were compared statistically. As a general rule, samples were classified as belonging to a particular group if they stated they were of that group, and if both their parents did so as well. For the family samples it was assumed that the cultural identity of the children was the same as that given by their parents, and for the samples from the Galton laboratory (which were used to investigate haplotypes rather than population groups) less detailed information was available, and so individuals were grouped by country of origin.

2.2 DNA Extraction

In most cases, DNA was extracted from buccal cells, though in some cases (in particular, the Finnish samples) blood was used. Buccal cells had been previously collected from volunteers using one of the two following methods.

i). A sterile applicator was rubbed over the epithelial cells around the inside of the mouth, focussing on the inside of both cheeks for approximately half a minute. The swabbed applicator was then placed inside a tube containing 1ml of 0.05M EDTA / 0.5% SDS preservative. On reaching the lab, the tubes were then stored at –20°C until extraction.

ii). The alternative method was as above, but ten cotton buds were used, and were stored in a tube containing 2.5ml of an alternative 'Slagboom' buffer. This was prepared as follows: 100ml 1M NaCl; 10ml 1M Tris HCl pH8.0; 20ml 0.5M EDTA pH8.0; 50ml 10% SDS, and distilled water was added to make a total volume of 990ml. 10 ml of 20mg/ml Proteinase *K* was added prior to use.

2.2.1 Extraction of the samples used for chapter 4

Most of the samples from TCGA and all of those used from the Galton laboratory had been extracted previously. The Ethiopian samples were extracted during the course of this PhD using the following methodology:

40µl of 10mgml –1 proteinase K was added to 20ml of sterile water and inverted repeatedly to mix. 0.8ml of this solution was added to each of the mouth swab sample tubes, which were then incubated at 60°C overnight. 0.8ml of each mouth swab solution was added to a separate, labelled microfuge tube containing 0.6ml of phenol/chloroform (1:1). The tubes were then mixed and centrifuged for 10 minutes at maximum speed (13,000 rpm) using a Hereus Biofuge Pico centrifuge. As much of the aqueous phase as possible was transferred to a sterile microfuge tube containing 0.6ml of chloroform and 30µl of 5 M NaCl. As before, the samples were mixed and centrifuged, and the aqueous phase was transferred to a new microfuge tube containing 0.7ml of chloroform, which were again mixed and centrifuged. The final aqueous phase was transferred to labelled screw-top tubes containing 0.7ml of isopropanol. These tubes were then inverted several times to mix and then placed in a freezer for a minimum of 2 hours.

After defrosting, the tubes were centrifuged as before for 12 minutes then the supernatant was discarded, and the tubes inverted at 45° to drain for 1 minute. 0.8ml of 70% ethanol was added to each of the sample tubes, which were then inverted and centrifuged again for 10 minutes. The supernatant was discarded

again, and the tubes were inverted at 45° to dry out for 20 minutes. 300µl of TE (10mM Trei-HCl, 1mM EDTA, pH 8.0) was added to each sample tube. The tubes were incubated in a water bath for 10 minutes at 56°C, and occasionally mixed during this time. The samples were then centrifuged briefly and stored at -20°C.

To confirm that the extraction had been successful 0.5µl of extracted genomic DNA sample was mixed with 6.5µl of H_20 and 3µl of dextran blue loading buffer (40% sucrose, 0.005% bromophenol, 0.05% xylene cyanol). The samples were then run on a 0.8% agarose gel at 100 volts per cm for approximately 20 minutes and visualised under UV light using ethidium bromide staining (4µl of EthBr 10mg/ml H_2O for a 100ml gel). Any remaining swab solution from each sample, and also the swab used to collect the DNA, was transferred to a labelled 1.5ml microfuge tube. These tubes were then sealed with Nesco film and stored at – 20°C as a back up sample.

2.3 Sequencing

The region of interest was amplified using the appropriate primers and PCR conditions (see section 2.4), and 2µl of the PCR product was mixed with a standard loading buffer and electrophoresed on an agarose gel to confirm the PCR was successful. The confirmation gel was also used to approximate the quantity of PCR product.

2.3.1 Purification of PCR product

MicroClean (Microzone ltd – 40% PEG-8000, 1 M NaCl, 2mM Tris-Hcl (pH 7.5), 0.2mM EDTA, 3.5 mM MgCl2) was diluted with distilled water (2 parts MicroClean: 1 part H_20). Next, an initial PCR template was purified by the addition of 3X volume of 2/3 MicroClean then mixed thoroughly, and centrifuged at 4500rpm for 40 minutes in an IEC Centra 4b plate centrifuge, all in the original PCR plate. The PCR lids were then removed, and the plate was

inverted onto tissue and spun at low speed (<200rpm) in the same centrifuge for 30 seconds to remove the supernatent. 150µl of 75% ethanol was then added to each sample and centrifuged at high speed (2000-4000g) for 20 minutes. The lids were then removed as before, and again the plates were inverted onto tissue and centrifuged at low speed (<200rpm) for 30 seconds. The tubes were left to air dry for 15 minutes at room temperature. 80µl of water was added to each sample, and then mixed thoroughly to disolve the DNA pellet. Six µl of this solution was subjected to electrophoresis on an agarose gel to confirm that it contained DNA, and to determine, as far as possible, DNA concentration. 5.5µl of template was then added to the sequencing PCR plate.

2.3.2 Sequencing reaction

A sequencing mix of the following reagents was prepared with amounts given per sample:

Better Buffer [Microzone Ltd – (200mM Tris-HCl pH9, 5mM MgCl2)] – 5µl, Termination mix [ABI prims BigDye Terminator kit] – 1µl
Primer 1.6 pmol/µl – 1.5µl
H_2O - 2µl
9.5µl of the mix was added to each DNA sample, making a total reaction volume of 15µl.

Cycle sequencing was performed using the following cycling parameters:

96°C for 10 seconds
50°C for 5 seconds ⎫ for 25 cycles
60°C for 4 minutes ⎭

4°C hold

2.3.3 Clean up of sequencing reaction products

The primers, unincorporated nucleotides / dye terminators and other reaction ingredients were removed from the sequencing products by adding to each reaction 80µl of 80% isopropanol, then mixing. The PCR plate containing the sequencing products mixed with isopropanol were then left for 10 minutes at room temperature. The PCR plate was then centrifuged for 40 minutes at 4500rpm, then inverted onto tissue and centrifuged again at low speed (<200rpm) for 30 seconds to remove the supernatent. 150µl of 70% ethanol was added to each sample to wash the DNA pellet, and the plate was centrifuged minutes at 4500rpm for 10 minutes. Then, to remove the supernatent, the PCR plate was inverted onto tissue and centrifuged again at low speed (<200rpm) for 30 seconds. The pellets were then left to air dry for 15 minutes at room temperature.

2.3.4 Electrophoresis of the sequencing products

The samples were prepared for electrophoresis, then subsequently run by a laboratory sequencing service as follows: 10µl of HiDi (high purity) formamide was added to each sample, which was then mixed and heated at 65°C for 5 minutes to ensure that the sequencing products were fully dissolved in the formamide. The PCR plate was then centrifuged at 1000rpm for 1 minute, and heated at 95°C for 4 minutes to denature the sequencing products. The products were then placed on ice for 5 minutes before being run on an ABI 3100 genetic analyser. The products were electrophoresed for 2.5 hours using POP6 polymer and capillaries.

2.4 Polymerase Chain Reaction

A series of polymorphisms in and around the lactase gene (see fig 2.1 below) were selected for typing. Each PCR was optimised for primer concentration, $MgCl_2$ concentration, annealing temperature and cycling conditions. The volume

of DNA template added varied between 0.5 – 2μl, depending on the estimated concentration of DNA in the sample used, but with an average amount of 20ng per reaction. Once optimised, all reagents were premixed in batches sufficient for a 96-well plate of 10μl reactions, and stored at –20°C. Taq polymerase was added prior to each reaction.

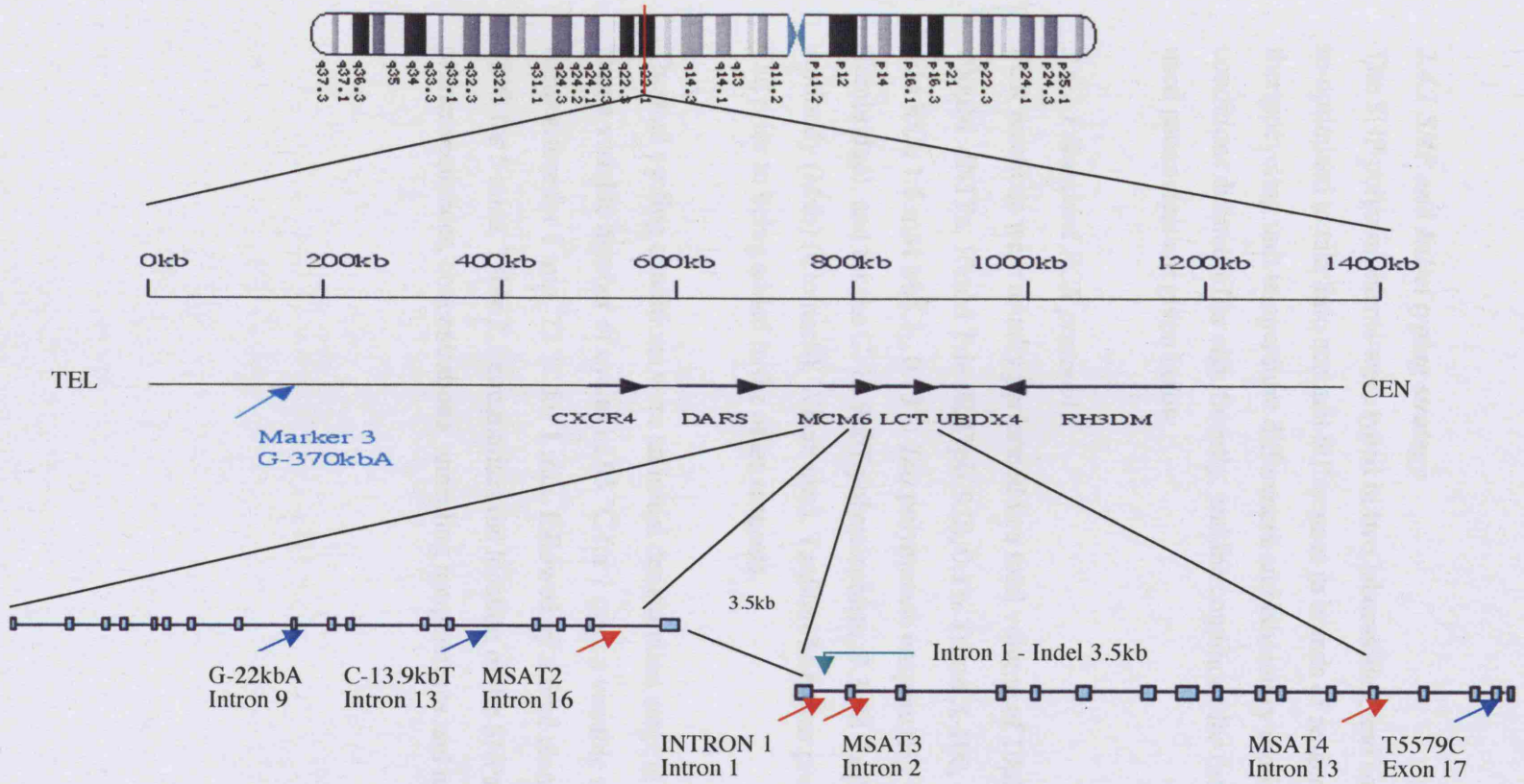


Fig 2.1 A diagram to show the polymorphisms investigated in this thesis. *Black arrows indicate genes and the direction of transcription, and lactase is abbreviated to LCT. SNP loci are indicated by red arrows, microsatellite loci with blue and an InDel polymorphism with a green arrow. Tel and Cen indicate the orientation of the region, 'telomeric' and 'centromeric' respectively. One of the microsatellites is not shown, D2S2385, which is located 1.2 MB in an intergene region upstream from the lactase gene*

2.4.1 SNP and InDel typing strategy

The SNP polymorphisms were typed in two laboratories, and some assays were re-optimised to take into account differences in brands of reagents, thermocycling unit temperature differences and laboratory routines. The reaction conditions did not differ significantly, and the conditions for the most frequently used parameters are given below.

2.4.2.1 Standard PCR protocol

PCR reactions were usually performed in a total volume of 10µl containing 200µM dNTPs, 10mM Tris-HCl (pH 9.0), 0.1% Triton X-100, 0.01% gelatin, 50 mM KCl, 1.5 mM $MgCl_2$, 0.13 U *Taq* polymerase enzyme (HT Biotech, Cambridge), and for the C–13.9kbT polymorphism, 9.3nM TaqStart Monoclonal Antibody (Mab) (Clontach). Where used, TaqStart Mab was premixed with the Taq prior to being added to the other reagents.

Thermal cycling conditions were an initial denaturation stage at 95 °C for 5 min, then a variable number of cycles of 95 °C for 1 min, a variable annealing temperature for 1 min, 72 °C for 1 min, followed by a final elongation stage at 72°C for 5 mins. Table 2.1 summarises the location of the SNPs typed, the primer sequences, concentrations, annealing temperatures and number of cycles.

Polymorphism	Primer Name	Primer Sequence 5'-3'	Concentration in reaction	Annealing temperature °C	Number of Cycling rounds
C-13.9kbT	LAC-C-M-U	GCTGGCAATACAGATAAGATAATGGA	0.3μM	55	35
C-13.9kbT	LAC-C-L	CTGCTTTGGTTGAAGCGAAGAT	0.3μM	55	35
G-22kbA	PEL 2 S	TTTGATGTTGGCTGATAATGTAAG	0.5μM	53	32
G-22kbA	PEL 2 A	CAATTTTAAATTCTAGATGAA GAAAAC	0.5μM	53	32
T5579C	X17 f	CTGAGAACTCAAATCAGCGC	0.25μM	56	30
T5579C	EXON 17 r	CACAGAAGCACAGACAGCTTT	0.25μM	56	30
Marker 3	P19p2f	TCATGGCAGCCTTAGTATATC	0.5μM	50	30
Marker 3	P19p2r	AGTCTGGATCCAGAAATCTG	0.5μM	50	30
Long allele – intron 1	Pdb24r	GTGGAATGTGAAACGGATCC	0.5μM	59	32
Long allele – intron 1	Pdb24	AGGACCATATGGCTGTCTTC	0.5μM	59	32
Long allele – intron 1	W15	GAAAACAGTGCAGTGCTACC	0.5μM	59	32
Long allele – intron 1	W1A	AGGTGTGTGATGAAGGTTGC	0.5μM	59	32
Short allele – intron 1	I1F3	CTAGGACATCATAGCTGCCT	0.5μM	59	32
Short allele – intron 1	F2S-REV	CTCTGACTGTGGAAACCACTG	0.5μM	59	32
Short allele – intron 1	5FS	GGAGGGTGAAGGAATTTGCAAG	0.5μM	59	32
Short allele – intron 1	PROA	GACTACATGCCAAGACAGCTCC	0.5μM	59	32

Table 2.1 Details of primer sequences used for SNP and InDel-intron1 analysis

2.4.1.2 Allele-specific enzyme digest conditions

Some of the SNPs typed were located at natural cut sites for a given enzyme, however, for C–13.9kbT, an engineered cut site was created using a modified PCR primer. In this case, the final or penultimate base of the upper primer introduced a base change such that the PCR product was cut by a restriction enzyme, *Hinf*1, and two different product sizes were generated depending on whether the allele of interest was present or absent (see table 2.2). Digestions were performed at 37 °C overnight in the original PCR plate in a total volume of

25µl for the –13.9kb C/T polymorphism (where each reaction contained the entire PCR product), and in 12µl reaction volumes for all other SNPs. The amount of amplicon used as a template ranged from 2µl – 4µl depending on the yield of the PCR, which was established by electrophoresis of 3µl of product mixed with 2µl of loading buffer on a 2% agarose gel for ten minutes. The table below summarises the enzyme and buffer used for each polymorphism, the number of units used, and the expected product sizes.

SNP	Enzyme	Buffer	No. of Units	Recognition sequence	Expected Product sizes
C-13.9 kbT	Hinf 1	NEB Buffer 2[5]	0.2	5'...GAN[6]TC...3' 3'...CTNAG...5'	177 and 24 (T) or 201 (C) bp
G-22 kbA	HinP1	NEB Buffer 2	0.4	5'...GCGC...3' 3'...CGCG...5'	124 and 353 (A) or 487 (G) bp
T5579C	Msp 1	NEB Buffer 2	0.4	5'...CCGG...3' 3'...GGCC...5'	109 and 34 (C) or 143 (T) bp
Marker 3	NlaIII	NEB Buffer 4[7]	0.4	5'...CATG...3' 3'...GTAC...5'	180 and 144 (A) or 324 (G)bp

Table 2.2 Details of enzymes used for SNP analysis

2.4.1.3 Electrophoresis of the PCR products

Ten µl of digestion product was mixed with 4µl of loading buffer and electrophoresed on a 2-3% agarose gel for 30-60 mins (depending on the expected product sizes). DNA bands were visualised using ethidium bromide staining to identify and interpret the specific digestion products and gel phenotypes, and a photographic record was kept.

[5] The composition of NEB Buffer 2 (1X) is: 10mM Tris-HCL, 10mM $MgCl_2$, 50mM NaCl, 1 mM dithiothreitol (pH 7.9 at 25°C)

[6] 'N' represents A, G, T or C, that is, the enzyme will recognise any of these bases in the sequence.

[7] The composition of NEB Buffer 4 (1X) is: 10mM Tris-acetate, 10mM magnesium acetate, 50mM potassium acetate, 1 mM dithiothreitol (pH 7.9 at 25°C)

For the agarose gels, various size standards or 'ladders' were used to identify the size of PCR products. That is, amplicons of known size (described for each gel on the figure legend) were used to compare the fragment size of amplified PCR products generated by each experiment. These ladders were made previously from DNA from a patient cohort panel of UK individuals. PCR was used to amplify a region of DNA from these samples, then restriction enzyme digest was used to produce a series of fragments of known size, ranging from 200bp to 1kb.

2.4.1.4 Genotype error checking

For each polymorphism investigated, one positive control of known genotype and one negative control was included in every 96 well plate, and all controls were consistent with the known status of the sample. For C-13.9kbT polymorphism, a set of 50 randomly selected samples were retyped blind and matched the initial typing. Where the results were ambiguous or unusual, the experiment was repeated.

2.4.2 Microsatellite typing strategy

A PCR 'multiplex kit' was designed to simultaneously amplify the regions of five selected microsatellites located in and near the lactase gene (see fig 2.1). Multiplex kits require extensive optimisation, as one amplification reaction can inhibit another. Preliminary experiments showed that amplification of intron 1 was inhibited by the other primer pairs, and so the intron 1 PCR was optimised separately as a 'singleplex' PCR.

Polymorphism	Primer Name	Primer Sequence 5'-3'	Concentration in reaction	Primer Label	Size range of product in base pairs	Observed number of repeat units
D2S2385	D2S2385F-F	CTGCTGACCTTTATCCACCTT	0.2μM			
D2S2385	D2S2385R	GTCCATAACCTTATAATGGTTC	0.2μM	TET	129-149	11-27
Microsatellite (2)	Msat2F	AAGATTTTCAACATTTGTATTTGAA	0.5μM			
Microsatellite (2)	Msat2R-F	TCATGTAGGCCTTTGTAGAGC	0.5μM	FAM	396-416	4-8
Intron 1	Int1-U	GCAAGACACAATGTGAAAAAAAAAAAA	0.2μM			
Intron 1	Int1-L	CCTCCAAGTCAGGGTAGCAGGACA	0.35μM	FAM	159-179	4-23
Microsatellite (3)	MS3F	GAGCACTATGGTGATGCATTC	0.35μM			
Microsatellite (3)	2MS3R	CTAAACCAAATATTCCAAAGCAG	0.35μM	HEX	146-156	9-19
Microsatellite (4)	2MS4F	TCAGAATTTGCATATGTTGTTTG	0.35μM			
Microsatellite (4)	2MS4R-F	CTACAAGAGCCTCTGAACCTG	0.35μM	FAM	123-131	9-18

Table 2.3 Details of primer sequences for microsatellite analysis

2.4.2.1 PCR Amplification of the kit

Primer pairs for four of the microsatellite polymorphisms were already designed (Hollox et al, unpublished), and the fifth primer, for the intron 1 polymorphism, was designed using Oligo v4.01 software. Primers were selected to minimize the possibility of false priming (especially at the 3' ends within the amplified region), and dimer formation, and also to have similar annealing temperatures. For each primer pair, one primer was labelled with an ABI fluorescent dye label (HEX, TET or FAM) selected such that overlapping size ranges were given different colours (see table 2.3). TaqStart Monoclonal Antibody was used to

increase the specificity of the PCR and was premixed with the Taq and stored at –20°C until needed (Thomas et al 1999). The antibody prevents the Taq from being active until it is destroyed by the first denaturation step, in effect producing a hot start reaction. The four sets of primers used in the kit were mixed and stored as a 10X stock concentration to minimize the likelihood of errors due to pipetting small volumes.

Reaction conditions were optimised for primer concentration, DMSO concentration (if required), $MgCl_2$ concentration and annealing temperature, and a range of DNA concentrations were tested to ensure sufficient sensitivity. PCR reactions were performed in 10µl volumes containing the following reagents:

200µM of dNTPs
Buffer: [10 mM of Tris HCl (pH 9.0) 1% Triton-X-100 0.01% gelatin, 50mM KCl]
0.13 units Taq polymerase enzyme (HT Biotech)
2.4 µM TaqStart Monoclonal antibody (Clontech)
Primers to concentration given in Table 2.3

For the main microsatellite 'kit', the optimal $MgCl_2$ concentration was found to be 2mM, and for the intron 1 PCR, which was typed separately, 1.5mM. The Intron 1 typing protocol included 2% DMSO in the reaction to increase specificity. Amplification reactions for both PCRs were performed in a Geneamp PCR system 9700 thermocycling unit (PE-Applied Biosystems) at the following conditions:

Pre-incubation at 95°C for 5 minutes

1 minute at 94°C
1 minute at 58°C — 35 cycles
1 minute at 72°C

72°C for 5 minutes

2.4.2.2 Acrylamide gel visualisation

For microsatellite analysis, it is important to size the DNA fragments reliably, and the ABI377/GeneScan (PE-Applied Biosystems, Foster City, CA) was used to resolve base pair differences in amplicons. This genotypying technique utilises fluorescently labelled primers, which, following PCR, are loaded onto a denaturing 5% polyacrylamide gel for electrophoresis. The laser detects the fluorescent signal of the labelled PCR product and the intensity of the signal (which depends on the number of labelled DNA molecules) is also recorded. The GeneScan system records the time taken from the start of electrophoresis for the product to migrate through the gel to the laser, and time is inversely related to the size of the fragment. The associated GeneScan software can be used to interpret the raw data files of fluorescent peaks.

The PCR products from the multiplex and the intron 1 'singleplex' reactions were mixed in equal volumes for each sample, then 1.1µl of the combined amplicons was added to a loading buffer. This consisted of an internal size marker (TAMRA 500, PE-Applied Biosystems), a dextran blue dye solution provided with the size marker and deionised formamide in a ratio of 2:4:24. The samples were then heated at 95°C to denature them and placed on ice until loading to prevent reanealing. The ABI-377 Automated Sequencer was used to electrophorese the PCR products, on a 36cm 5% acrylamide gel for 3 hours.

2.4.2.3 Genescan analysis of allele-specific size fragments

The Genescan system stores information from the laser reading as a 'gel file' that can later be extracted using ABI PRISM collection software. The gel file of raw data was then analysed using Genescan Analysis v3.1 software. The relative size of PCR amplicons was determined as part of the program by comparison with the size standard, which contained DNA fragments of the following sizes: 50; 75; 100; 139; 150; 160; 200; 300; 340; 350; 400; 450; 490; 500.

Having recorded the estimated size in base pairs of a given PCR amplicon (in relation to the size marker) to one decimal place, the next stage was to calculate the number of tandem repeats for each microsatellite. Initially, a sequence from the genome browser[8] of known size and tandem repeat number was used to estimate the number of microsatellite repeats for a PCR product of a given size. For each of the loci, the estimated sizes of all the samples were taken from the Genescan output (raw data) and graphs of frequency were produced in excel. These were used to estimate the most probable number of repeats for each series of clustered results, which were assumed to be alleles. Once the repeat numbers had been established, a margin of error was set at 2/3 of the number of nucleotides in the repeat motif sequence for each microsatellite. This was done to account for minor variations in the gel run. A formula was then created in Excel such that any result recorded which fell outside of this range could be labelled 'bad data', and then the PCR product could be rerun on another gel at a later stage. This microsatellite repeat number was recorded in a separate table for analysis.

[8] The Human genome browser found at: http://genome.ucsc.edu/

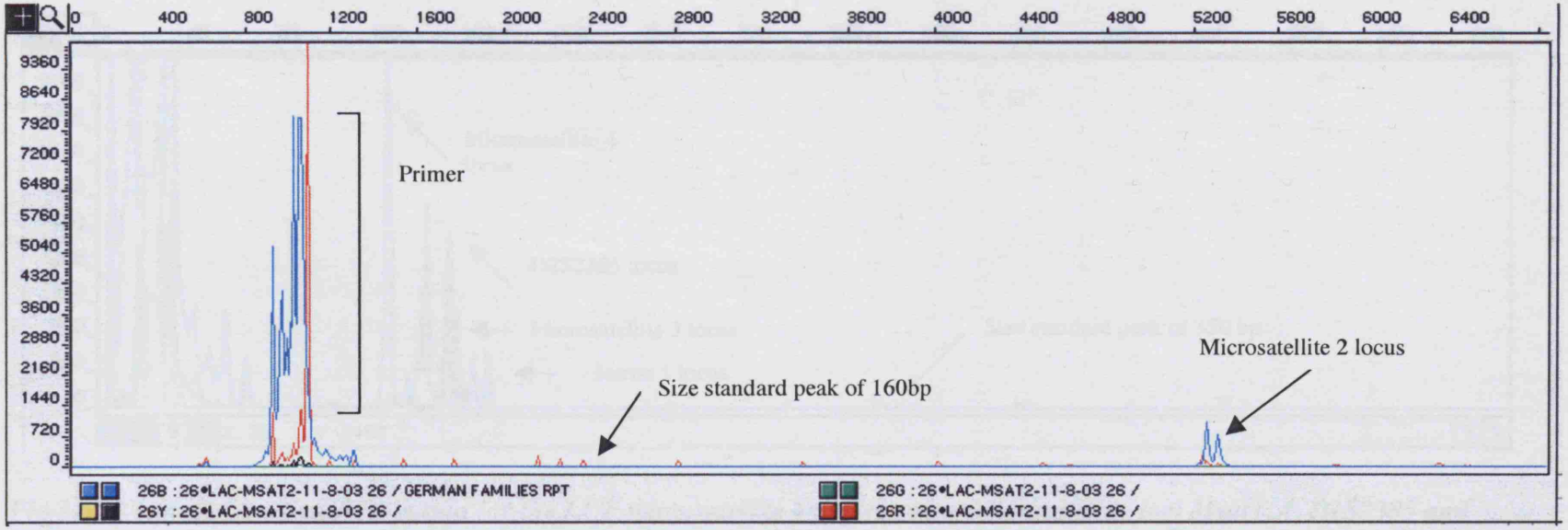


Fig 2.2. A typical GeneScanTM output for the LCT microsatellite assay showing the Microsataellite 2 locus, which generated a product size larger (200bp approx.) than the other four loci. *The peaks shown represent fluorescence detected by the ABI 377 laser, which records a signal for each detected peak proportionate to the time taken for the amplicon to migrate through the acrylamide gel. Here, a heterozygote is shown for msat2. The smaller red peaks represent the internal size standard as described in 2.4.2.3.*

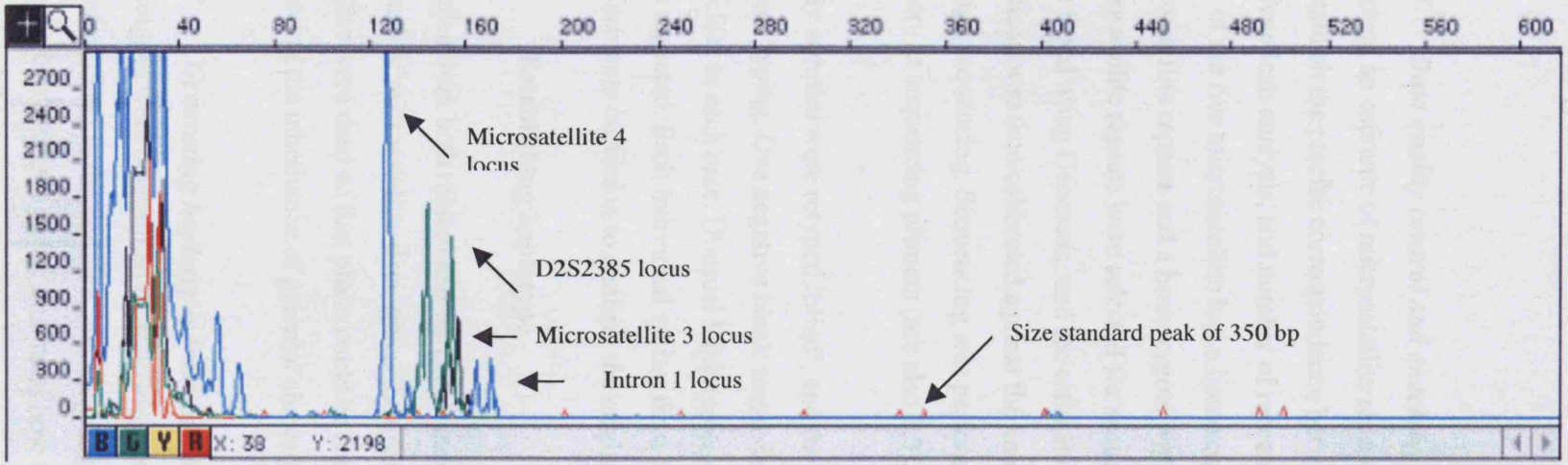


Fig 2.3. A typical GeneScanTM output for the LCT microsatellite assay showing Microsatellite loci Msat3, 4, D2S2385 and Intron1.

The peaks shown represent fluorescence detected by the ABI 377 laser, which records a signal for each detected peak proportionate to the time taken for the amplicon to migrate through the acrylamide gel. Here, the first microsatellite peak is a homozygote for msat4, the two green peaks, a heterozygote for D2S2385, the black peak is msat3 and the two smaller blue peaks are intron 1. The smaller red peaks represent the internal size standard as described in 2.4.2.3.

2.4.3 Data quality control and checking

Although an estimate of microsatellite repeat number was used, it was necessary to establish the precise correcpondence between PCR product size, as determined by GeneScan analysis, and number of repeats of each microsatellite motif. For each of the five microsatellite loci a homozygote with the lowest number of microsatellite repeats and a homozygote with the highest number of microsatellite repeats were selected for sequencing. These samples had already been typed using Genescan, and the estimated number of repeats for a given amplicon was then calibrated against the known number of repeats as determined through sequencing. Sequencing was performed in both directions using PCR primers as sequencing primers (see also 2.3).

Thirty samples were retyped 'blind', and the results obtained matched the original typing. One negative blank was included for each PCR plate, and this was clear in each case. Unusual haplotypes, or readings that had been ambiguous were repeated. Each individual typing from the Genescan was re-analysed blind on a separate occassion to confirm the original reading.

2.5 Establishing haplotype

Samples from both related and unrelated individuals were used in this study. In the case of microsatellite data, given the greater allelic diversity, populations of families were used so that phase could be accurately determined through observing the inheritance of parental chromosomes in one or more child.

2.5.1 Generating haplotypes from families

Haplotypes were determined by following the pattern of inheritance (Nejati-Javaremi 1996) in the families, and assuming no recombination. Figure 2.2, an example of a pedigree sheet, illustrates how the genotypes of both parents and at least one child were resolved.

Analysis of a series of families from HapMap SNP data described in chapter 7, the number of polymorphic sites for which genotypic data was downloaded was too great to use this method, and so the PHAMILY program was used to resolve phase and determine haplotypes.

2.5.2 False paternity in family samples

In cases of apparent non-Mendelian inheritance pattern, there are two possible interpretations, the most likely of which is incorrect typing. Checking the data entry or repetition of the experiment usually resolves this type of error. However, when multiple alleles are found in children and not observed in either of the parents, there is a possibility of false paternity that may, if unidentified, lead to incorrect haplotyping. The families in this thesis have been examined at a number of other loci, and several cases of possible non-paternity were identified (Fletcher 2002, Caldwell 2005). Out of 147 families, 7 in total did not conform to a normal pattern of Mendelian inheritance and were later discarded on the basis of non-paternity[9]. These results corresponded with the previous findings, except in 1 case, which would not have been recognised using the lactase haplotype data alone, since the both parents and children had one of the more common haplotypes, and, due to higher levels of homozygosity in the population, the inheritance appeared unproblematic. Chromosomes from these families were not included in the analysis.

The frequency of non-paternity ranged from 0% in the Irish, French CEPH, Armenian, and Algerian families, to 10 % approx. for the German and English families.

[9] It should be noted however that mislabelling of the mouthswab samples at the time of collection has not been excluded

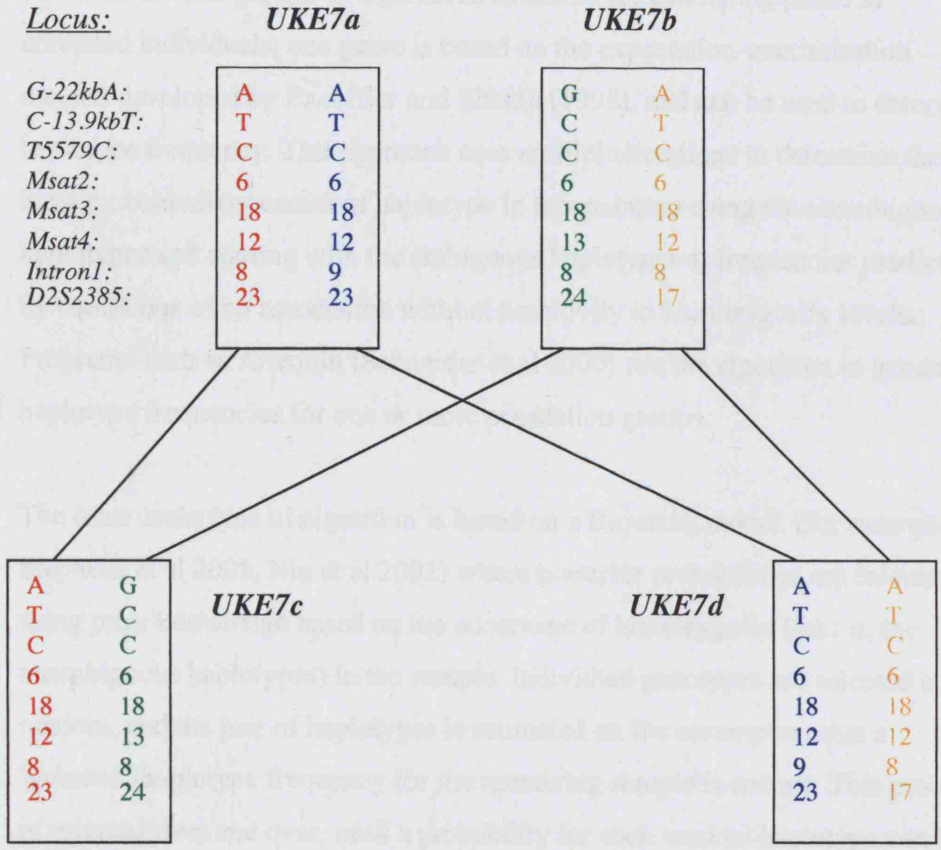


Fig 2.2 A typical pedigree to illustrate haplotyping method

This pedigree was taken from a UK English family (no.7)
In this example, loci Msat2 and Msat 3 are homozygote for both parents, and Intron 1, Msat 4 and D2S2385 are homozygote for one parent. These alleles can therefore be assigned without complication. In child 7c, alleles from loci Msat 4 and D2S2385 can only come from parent 7b, and so can be demonstrated to exist on the same chromosome. Other alleles were assigned using the same approach.

2.5.3 Establishing Haplotypes in unrelated individuals

There are several groups of algorithms available for estimating phase in unrelated individuals; one genre is based on the expectation-maximisation method developed by Excoffier and Slatkin (1995), and can be used to determine haplotype frequency. This approach uses multiple iterations to determine the most probable frequencies of haplotype in a population using the unambiguous haplotypes and starting with the ambiguous haplotypes at frequencies predicted by conditions of no association without sensitivity to homozygosity levels. Programs such as Arlequin (Schneider et al 2000) use the algorithm to generate haplotype frequencies for one or more population groups.

The other main type of algorithm is based on a Bayesian model, (for example, Stephens et al 2001, Niu et al 2002) where posterior probabilities are calculated using prior knowledge based on the occurance of homozygotes (that is, the unambiguous haplotypes) in the sample. Individual genotypes are selected at random, and the pair of haplotypes is estimated on the assumption that a 'guessed' haplotype frequency for the remaining sample is correct. This process is repeated over and over, until a probability for each worked haplotype can be established. This error rate can be useful for gauging the reliability of the estimated haplotyped frequencies for any given sample group to be used in analysis. Frequencies of haplotypes are generated, but also, in many cases, the individual haplotypes from each individual are estimated.

For the unrelated individuals examined in this thesis, the PHASE program was used[10]. PHAMILY, a similar program for use with families was also used to resolve cases in the families where heterozygosity at multiple loci made it difficult to determine segregation. The program estimates the haplotypes in

[10] Available from web-site http://archimedes.well.ox.ac.uk/pise/phamily-simple.html

problematic cases by observing frequencies of resolved haplotypes in the same population.

2.6 Statistical procedures

For many investigations in this thesis, a null hypothesis was generated and tested using a variety of statistical techniques. In all cases, p-values less than or equal to 0.05 were considered 'significant', p-values less than or equal to 0.01 were considered 'highly significant', and both levels were used as evidence that the null hypothesis could be rejected.

2.6.1 *Hardy-Weinberg*

The Hardy-Weinberg equation, $p^2 + 2pq + q^2 = 1$, summarises the expected proportion of genotypes for a biallelic locus in a population group, where allele frequencies are known. p^2 and q^2 refer to the homozygote frequencies and 2pq is the frequency of heterozygotes. These proportions are expected for a population in equilibrium, and assume conditions of random mating, no selection and no population sub-structuring. Deviations from Hardy-Weinberg can occur if these assumptions are not met, or if there are null alleles, or if there are technical problems with the data set such as mistyping or selective allele drop-out. Arlequin was used to determine Hardy-Weinberg equilibrium, based on the method of Guo et al (1992).

2.6.2 *Fishers Exact test and Chi Square test (χ^2)*

Some comparisons between anthropological data and genotype frequency data were made using either Fisher's exact test (for example, Ott 1991) or the χ^2 test (for example, Fisher et al 1938). Fisher's exact test was used in cases of a two-by-two contingency table. A requirement of using the test is that two categorical variables with only two possible states are used. The table generated by the data is compared with the expected table if there were no association. Then, all the possible tables that deviate from the expected distribution further than the

observed table in both directions are calculated, first in the direction of deviation of the observed data set, then in the other direction, which includes probabilities exceeding the observed data set. From this, a two-sided p-value is obtained (Agresti 1992).

Where more than categorical variables with more than two states were under investigation, a χ^2 test was used. The formula for this was used as follows:

$$\chi^2 = \frac{\Sigma (observed/expected)^2}{expected}$$

Where the total of any of the values under investigation was under 5, -0.5 was subtracted from each value (Yate's correction).

A web-based statistical calculator[11] was used to calculate both Fisher's Exact and also χ^2.

2.6.3 Linkage disequilibrium (LD)

LD can be measured in several ways, (for review, see Devlin and Risch 1995, LD paper). Currently, the most commonly used measures are D', and 'r^2'. D' is the result of ***D/Dmax***, where D evaluates the deviation from the allele frequencies observed and those expected if there is no association, and Dmax is the maximum deviation possible given the allelic frequencies observed. A value of 1 indicates the maximum deviation from the expected distribution, that is, the maximum value of linkage disequilibrium that can be observed (Lewontin 1964). D' is limited as a statistic because where the numbers are low, a figure of 1 may not be statistically significant. However, D' can be used for multiple alleles.

The statistic r^2 (Hill et al 1994) uses a correlation coefficient between two loci squared, and is less sensitive to population size or a smaller frequency of the rarer allele.. Wherever possible, r^2 was used as it is a more robust measurement

[11] This was available at website: www.matforsk.no/loa/fisher.htm

against variation between population groups. r^2 can be summarised by the equation:

$$r^2 = D^2 / (ABab)^{0.5}$$

Calculations were performed using using HaploXT and Arlequin (Schneider et al 2000). Block-like patterns of LD were visualised in the Gold program, which creates a graphical representation of the blocks of linkage disequilibrium. The χ^2 test was also generated by HaploXT and shown in the Gold program.

2.6.4 Descriptive statistics

For the microsatellite haplotypes, tests of population differentiation were used to describe the genetic variation within and between the populations under investigation. Basic measurements of mode, median, mean and variance were calculated in Excel. Several standard tests were used to define and describe the degree of heterozygosity observed for specific loci in sampled population groups.

2.6.4.1 F_{st}, R_{st} and Analysis of Molecular Variance (AMOVA)

F_{st} describes the probability of two randomly sampled alleles being identical by descent within and between two (or more) population groups. F_{st} can be calculated by randomly sampling again from within a population and scoring '1' or '0' depending on whether they are identical or different (Hw), then sampling two chromosomes again, but using all population sets (Ht). The first procedure, if repeated, gives an estimate of the variance within a population set, the second gives the total variance, and so $F_{st} = (Ht - Hw) / Ht$. In order to get closer to a 'true' value of F_{st}, the sampling procedure has to be repeated over many iterations to generate F_{st} values. F_{st} itself is a less robust indicator of variance between groups where 'N' is different for the population groups under study, as sampling error can lead to bias, and this is also the case where the original populations sampled from are significantly different in size.

R_{st} is similar in principle to F_{st}, but is applicable only to microsatellites, and takes into accound allelic difference in repeat number, assuming a stepwise

mutation process, that is, that mutations usually result in a chage in repeat number by one unit (Di Rienzo et al 1994). Given this assumption, the frequencies of alleles can be ranked in order of motif repeat number and statistical tests used to compare these frequencies, such as a paired 'T – test' if two populations are being considered, or an analysis of variance (AMOVA) if more. F_{st}, R_{st} and AMOVA were calculated in Arlequin (Schneider et al 2000).

2.6.4.2 Exact test of population differentiation

The exact test of population differentiation (Goudet et al 1996, Raymond et al 1995) between pairs of populations was performed for microsatellite haplotype data. The statistic is analogous to Fisher's exact test and compares the observed distribution of alleles against an expected, random distribution. A contingency table is generated based on the number of haplotypes and the number of populations. An algorithm run in the Arlequin program (Schneider et al 2000) adds the probability of the observed table with that of all the possible tables that are less probable. A Markov Chain generates the different possible tables, and standard deviation from the mean is taken into account.

2.6.5 Comparison between frequency of an allele and recorded levels of lactase persistence phenotype

A statistical procedure designed by Dr. Mike Weale was used to compare frequencies of lactase persistence phenotype recorded from previous studies with the frequencies of a proposed causative allele observed in equivalent population groups as measured during the course of this thesis. The programme is available at the TCGA web-site and uses a comparison of observed and expected lactase persistence levels calculated by the presence of the –13.9kb*T allele.

The comparison accounts for sampling error for both the phenotyped and genotyped populations, as well as errors associated with phenotyping (both false negatives (*fn*) and false postitives (*fp*)), and is described in more detail in Appendix A1. The combined results from the above five studies enable the

program to perform a rough averaging over the differences in protocols used (blood glucose *fn* = 10/116 and *fp* = 5/73, breath hydrogen *fn* = 9/132 and *fp* = 5/120). The combined results suggest that even if a population has no lactase persistent individuals, we would expect, using one of these two methods, to find between 5% - 10% false negatives (i.e. apparent lactose digesters).

The statistical procedure devised by Mike Weale (2004) makes the assumption that the underlying phenotyping error rates acting in the studies reviewed are applicable to other studies that use the same measurement techniques, and can be used to investigate whether the frequency of lactose digesters predicted by the *C–13.9kbT* genotype data was sufficient to explain the observed frequency found in the phenotyped group. Four possible sources of sampling uncertainty were taken into account: (1) sampling uncertainty in *p*, the frequency of the *–13.9kb*T* in the genotyped group; (2) sampling uncertainty in *fn*, the frequency of false negatives according to the phenotyping method used; (3) sampling uncertainty in *fp*, the frequency of false positives according to the phenotyping method used; and (4) sampling uncertainty in *Lapp*, the frequency of apparent lactase persistence in the phenotyped group. A description of the calculation of the error rate can be found in Appendix A1.

2.6.6 Intra-allelic diversity based test for selection

The Syssiphos program, written by Dr. Michel Stumpf, utilises compound microsatellite and SNP haplotypes to assess intra-allelic diversity and was used in chapter 6 to investigate evidence of selection favouring the *13.9kb*T* allele. Multiple simulations are run to generate likelihoods of the data (allele frequency for the SNP under investigation and the microsatellite haplotypes occurring on that SNP background in the sample) given different values of selection and population growth. All the chromosomes for a given population sample are pooled, and two text-format input files are generated, *msats0* and *msats.in*. The first contains the microsatellite haplotypes for the population group as a whole, including all the microsatellite haplotypes associated with each possible SNP

allele. The second contains the microsatellite haplotypes for only one of the SNP haplotypes, which will then be run in the simulation. Different *msats.in* files enable simulations of each SNP haplotype. Simulations were not run for SNP haplotypes containing fewer than 8 chromosomes in the data set. The data file containing all microsatellite haplotypes present in the population is used to model the effects of recombination on intra-allelic variability and assumes that the current distribution of microsatellite haplotypes is the same as ancestral distributions.

Another input file enables various parameters to be set, some of which were investigated in chapter 6. Of these, the parameter most likely to be affected by changes in setting is mutation rate. Mutation rates are known to vary for microsatellites of different length, different repeat motif, sex, age and genomic region (Ellegren 2000). Some of these variables could be accounted for: an increase in microsatellite mutation rate with greater length of allele was accounted for by taking the length of the shortest allele, and using this to create a length dependent mutation rate; the intercept point for this was given as a fixed value of –0.62 (Stumpf, personal communication). A stepwise model of mutation was otherwise assumed. It was not possible to modify the mutation rate to take into account sex and age, or genomic region, so averaged rates were used (Weber and Wong 1993). The parameters used in Syssiphos can be summarised as follows:

Input code	Description of parameters
Nrun	Number of runs of program (set at 4000)
Tmax	Depth of coalescence, deepest possible root in generations (set at 100,000, with a generation time of 20 years)
I	Number of chromosomes of the particular SNP haplotype under analysis population
xT	Frequency of SNP haplotype in the total population group
nrmsat	Number of microsatellite loci
ne0	Effective population size – 1e7
mu	Mutation rate – 0.0012 (Weber and Wong 2003)
slow / shigh / dels	Maximum, minimum and integer values for selection – varied for each simulation, initially set as 0.0001, 0.1 and 0.05
rlow / rhigh/ delr	Maximum, minimum and integer values for growth – varied for each simulation, initially set as 0.0001, 0.1 and 0.05
rho	Recombination rates (as shown in chapter 6, 6.2)
A	Length dependency of microsatellite mutation rate – defined as mu (k) = mu (a +b) where k = allele length of the shortest observed allele, and a is equivalent to 'slope'. In this case slope set at -3.1 (see Stumpf and Goldstein 2001)
B	Length dependency of microsatellite mutation rate, defined here as mu (k) = mu (a + b) where b is the intercept, set here at - 0.62. (see Stumpf and Goldstein 2001)
msats0	Number of total chromosomes in population

2.4 Table of Syssiphos parameters

2.6.7 Dating demographic events – Ytime

The statistical procedure 'Ytime'[12], written by Dr. M. Weale (v2.05, online in 2004) was used to estimate coalescence dates of haplotypes carrying specific SNP alleles using microsatellite data (see Behar et al 2003). Ytime was originally intended for use with Y Chromosome clades, but can be used for a diploid system where phase is known if recombination is assumed to be abscent. Ytime is comprised of a series of functions for input into a MATLAB programming environment, and analyses data for a given allele, where data from linked microsatellite loci is known. The program calculates the Average Squared Distance (ASD) between the ancestral haplotype (assumed here to be the modal haplotype) and a sample of chromosomes (Slatkin 1995). ASD has been shown to be linearly related to the time since divergence, in generations, assuming a constant microsatellite mutation rate. As for the Syssiphos simulations, microsatellite mutation rate was assumed to be 0.0012 (Weber and Wong 1993), and a length-dependent model was used (Calabrese et al 2001, Kruglyak et al 1993). Population growth rate was set at zero.

[12] The program can be downloaded from the TCGA website at www.tcga.ucl.ac.uk

2.7 Manufacturers and suppliers

Sigma-Aldrich, St.Louis, Missouri (General)

Fisher Scientific, Loughborough, Leicestershire, UK (General)

Fissons Scientific Equipment, Loughborough, Leicestershire, UK (General)

Merck BDH Chemicals, Poole, Dorset, UK (General)

Sartstedt, Numbrecht, Germany, (Swab tubes for samples)

New England Biolabs, Beverly, MA (Restriction enzymes)

MWG Biotech Ebersberg, Germany (General)

Sigma Genosys (Oligonucleotides)

HT Biotech, Cambridge, UK (Taq Polymerase)

BD Biosciences Clontach, San Jose, CA (Mab)

Whatman Biometra, Goettingen, Germany (PCR machine)

National Diagnostics, Atlanta, Georgia (Acrylamide solution)

ABGene, Epsom, Surrey (dNTPs, PCR buffers)

Microzone Ltd, Haywards Heath, UK (sequencing reagents)

PE-Applied Biosystems, Foster Cuty CA (PCR Machine, DNA Sequencing and GenescanTM Equipment)

Chapter Three

Haplotypic background of alleles associated with lactase persistence phenotype

3.1 Introduction

As discussed in 1.1.6.3, two SNPs located in the *MCM6* gene were reported as having a very strong association with lactase persistence phenotype (Enattah et al 2002). The –13.9kb*T allele in particular showed a complete association with lactase persistence phenotype, and on the strength of this, was considered as a candidate for a causative mutation for the trait. Given this, an early priority was to repeat the findings of Enattah and colleagues (2002) and demonstrate this association in another population. This chapter investigates the association of these alleles with the core lactase haplotypes described by Harvey et al (1998) and Hollox et al (2001). It also characterises an Insertion / Delection polymorphism with respect both to core lactase haplotypes and the two reported alleles.

3.1.1 Sequencing panel of known phenotype and haplotype

The first stage in this study was to design an effective assay for typing the putatively causal –13.9kb*T SNP described by Enattah et al (2002), and to confirm its reported association with the lactase persistence phenotype. Following amplification of a DNA template from a panel of five UK individuals, (hereafter referred to as the 'UK Panel') the region of interest was sequenced in both directions (see section 2.3). Although the quality of the sequenced fragments varied, it was possible to resolve 177bp of good sequence data from each individual. The sequence chromatograms (fig 3.1 and fig 3.2) show samples 198, a CT heterozygote for the –13.9kb locus and sample 187, a –13.9kb*C homozygote.

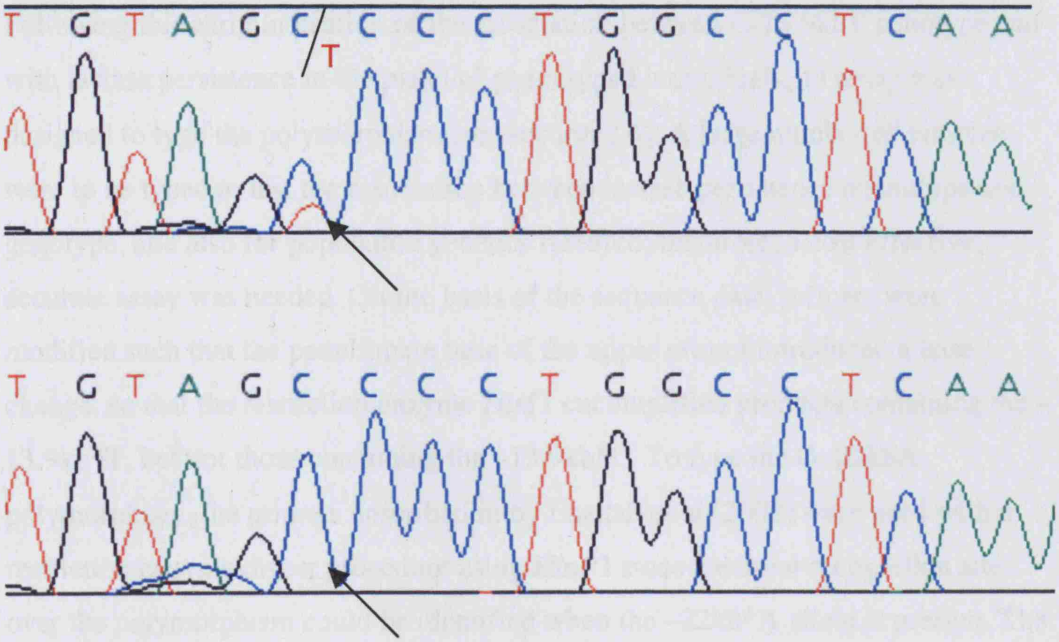


Fig 3.1 and 3.2 Sequence Chromatograms showing the C-13.9kbT loci.
The polymorphic locus is marked with an arrow.
*The upper chromatogram shows sample 198, a –13.9kb*CT heterozygote, and the lower chromatogram shows sample 187, a –13.9kb*C homozygote*

Phenotype for the UK panel had previously been established using jejunal biopsy samples and enzyme assay (Ho et al 1982), and the samples had also previously been typed for the extended core lactase haplotypes[13] (Harvey et al 1998, Hollox et al 2001). The –13.9kb*T allele was only found in the three lactase persistent individuals (see table 3.1), each of whom carried chromosomes typed as haplotype **AB**.

Sample name	Origin	Haplotype	Lactase persistence phenotype	C-13.9kbT Typing
182	London	AB	Persistent	C/T
187	London	BC	Non-persistent	C/C
198	London	AB	Persistent	C/T
209	London	AB	Non-persistent	C/C
210	London	AB	Persistent	C/T

Table 3.1 Extended core lactase haplotypes, lactase persistence phenotypes and genotypes for a panel of UK individuals

[13] Several papers have grouped SNPs to form different lactase haplotypes; the haplotypes first identified by Harvey et al (1998) and, later, further characterised by Hollox et al (2001) are described in this thesis as 'core lactase haplotypes' and shown in full in appendix B1

Following this early indication of the association between C-13.9kbT genotype and with lactase persistence in the panel of phenotyped individuals, an assay was designed to type the polymorphism (see section 2.4). A large number of samples were to be typed to test the association between lactase persistence phenotype and genotype, and also for population genetics research; therefore, a cost effective, accurate assay was needed. On the basis of the sequence data, primers were modified such that the penultimate base of the upper primer introduced a base change, so that the restriction enzyme *Hin*f1 cut amplified products containing the –13.9kb*T, but not those containing the –13.9kb*C. To type the G–22kbA polymorphism, the primers described in by Enattah et al (2002) were used with a restriction enzyme digest procedure using *HinP*1 since a natural recognition site over the polymorphism could be identified when the –22kb*A allele is present. The PCR and enzyme specific digest products were then visualised on agarose gels (see figs 3.3 and 3.4), and the genotypes were inferred by looking at the gel phenotypes.

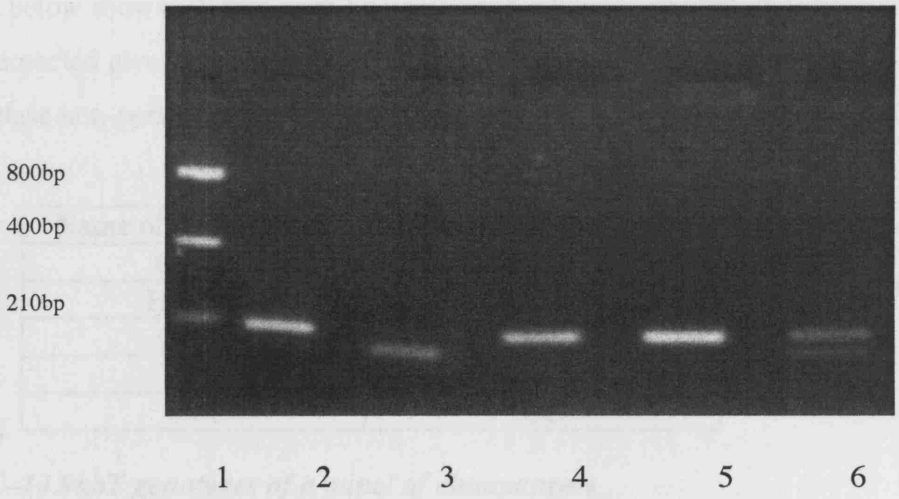


Fig 3.3 An example gel for typing the C-13.9kbT polymorphism
*Lane 1 shows a ladder, lanes 2-5 show a series of samples, where –13.9kb*T gel phenotype band is 177bp (a second band of 24bp is not visible) and the –13.9kb*C is 201bp. Lane 6 shows the positive control, a heterozygote (sample 182)*

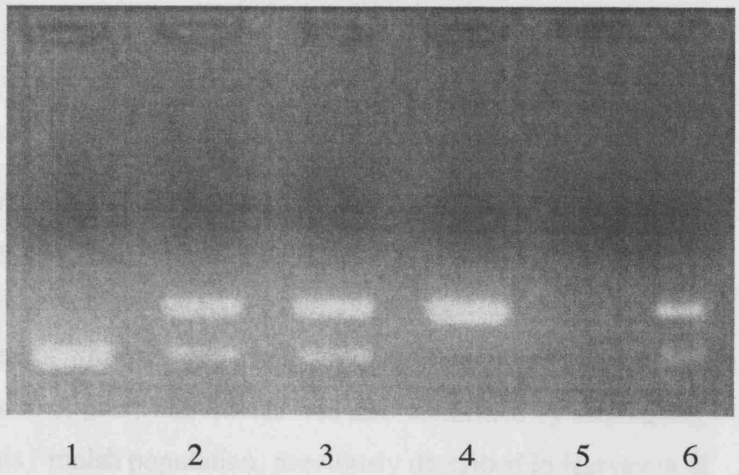


Fig 3.4 An example gel image for typing the G-22kbA polymorphism
*Lanes 1-4 show a series of samples, where –22kb*A gel phenotype band is 353bp and the –22kb*G is 405bp. Lane 5 is the negative control, and lane 6 shows the heterozygous positive control (sample 182)*

A panel of five chimpanzees from West Africa were then tested for the polymorphism to determine the most likely ancestral state at the C–13.9kbT site. As table 3.2 below shows, all five were homozygote for the –13.9kb*C allele, which was expected given the reported association between –13.9kb*C and the ancestral lactase non-persistent phenotype.

Name of chimpanzee	C-13.9kb*T genotype
Casey	CC
Harvey	CC
Colin	CC
Tank	CC
Carl	CC

Table 3.2 C–13.9kbT genotypes of a panel of chimpanzees
The panel were Pan troglodytes from West Africa

3.2 The association between –13.9kb*T allele and lactase persistence

Using the PCR and enzyme digest assay described in 3.1 and 2.4, members of the Galton Laboratory group tested two sets of phenotyped samples (Poulter et al 2003). 40/41 unrelated Finnish individuals showed a tight correlation between lactase persistence status and presence of –13.9kb*T and –22kb*A alleles. The C-13.9kbT genotype for three of these individuals was also confirmed by sequencing (data not shown). For this Finnish population, previously described in Harvey et al (1998), three indirect tests had been used: the breath hydrogen test, the blood glucose test and the urine galactose test. The technique of using three indirect methods of diagnosis, and taking a 'best of three' approach, is commonly known as 'the Gold Standard' (for example, Peukhuri 2000). Although it is noteworthy that there was one exception, these results in general supported the association reported by Enattah and colleagues (2002) between the 13.9kb*T allele and lactase persistence. Similarly, with only two exceptions, the –22kb*A allele was found in lactase persistent individuals but not in non-persistent individuals.

The association between –13.9kb*T allele and lactase persistence was further investigated using a second data set intestinal biopsy samples that had been obtained from 48 London patients of mixed ancestry. In this group, data on sucrase/lactase ratio had been previously established using an enzyme assay technique to determine lactase persistence phenotype, and also whether the individual was likely to be homozygote or heterozygote. The –13.9kb*T allele was present in all 36 lactase persistent individuals and not in 11 non-persistent individuals. However, for individuals that were heterozygous for the C-13.9kbT polymorphism, the results were more problematic. Several of the heterozygous individuals did not show intermediate lactase enzyme activity, as might be expected if the –13.9kb*T allele was causative of lactase persistence (Poulter et al 2003). In particular, sample 182, sequenced as part of the UK Panel, showed high levels of lactase expression consistent with homozygote status and expression of two allelic lactase mRNA transcripts at high level (Wang et al 2005). However, as the

sequence chromatogram below (fig 3.5) shows, the sample was a –13.9kb*CT heterozygote[14].

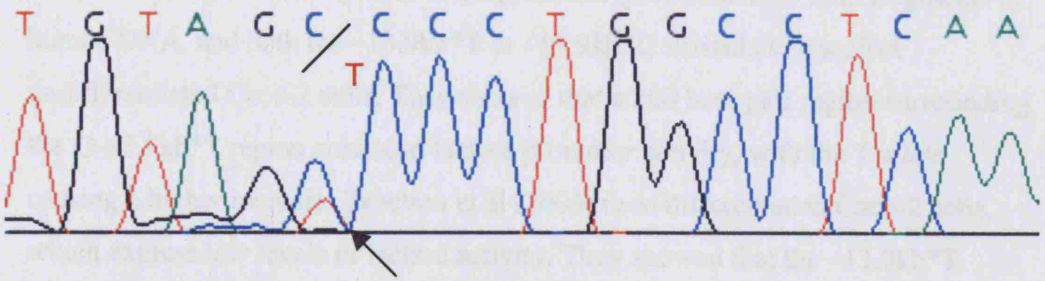


Fig 3.5 Sequence chromatogram showing sample 182
*The arrow indicates the heterozygote –13.9kb*CT site*

More recently, a study of 123 Austrian patients with suspected lactase non-persistence showed that there was an excellent correlation (97%) between individuals CC for the –13.9kbT polymorphism and a breath hydrogen test result of non-persistence. The same study showed that there was a less tight correlation between individuals with CT or TT genotypes and a breath hydrogen test result of persistence (86%) (Hogenauer et al 2005). Given the data from the intestinal biopsy samples, it may be the case that although the –13.9kb*T allele is a good predictor of phenotype in general, it is less sensitive in distinguishing between homozygotes and heterozygotes. However, another study by Kuokkanen et al (2003) suggested a strong association between *LCT* mRNA expression, lactase activity in the gut and C-13.9kbT genotype in a Finnish group. It may be the case that the –13.9kb*T allele associates completely with lactase persistence in Finland, but that this association is not complete elsewhere.

Recent studies suggest that the –13.9kb*T SNP is located in an enhancer element, and may explain observed functional differences between alleles carrying the –13.9*T and –13.9*C alleles (Olds and Sibley 2003, Troelsen et al 2003). Both studies showed that, when the two variants were cloned and used to transfect

[14] The sequencing for this sample was repeated three times

human intestinal Caco-2 cells, lactase promotor activity was differentially affected. Olds and Sibley used a sequence of rat promotor gene combined with fragments of human DNA, and both the –13.9kb*T or –13.9kb*C variants to transfect undifferentiated Caco-2 cells. They showed that a 200 base pair region surrounding the C–13.9kb*T region enhanced lactase promotor activity, with the T allele causing a higher increase. Troelsen et al (2003) used differentiated Caco-2 cells, which express low levels of lactase activity. They showed that the –13.9kb*T variant enhanced lactase promotor activity approximately four times as much as the –13.9kb*C. The authors provided evidence of a nuclear factor binding more strongly to the –13.9kb*T variant, strongly consistent with a causative mutation.

3.3 Establishing Haplotypic background of the –22kb*A and –13.9kb*T alleles

Data from the samples of unrelated Finns and also the 'UK panel' described gave an early indication that the –13.9kb*T and –22kb*A alleles were found uniquely on the background of the A Haplotype. Several sets of samples were used to investigate further the haplotypic background of –13.9kb*T and –22kb*A alleles.

3.3.1 CEPH families

A series of families collected by the Centre d'Étude de Polymorphism Humain (CEPH) were used to determine the haplotypic background of the –22kb*A and –13.9kb*T alleles. The CEPH families used were collected from various regions in Northern France (Dausset et al 1990). Information from a series of polymorphic markers that had been used to identify five core lactase haplotypes was already available (Harvey et al 1998). Phase was established using the CEPH family pedigrees[15] (Nejati-Javaremi and Smith 1996), and children from each family were used to resolve parental chromosomes. A series of CEPH French families were typed in total, enabling analysis of 48 parental chromosomes. The pedigree analysis from this data set suggested that both the –22kb*A and –13.9kb*T alleles described

[15] Pedigree information is available at website: http://landru.cephb.fr/

by Enattah et al (2002) were inherited on the background of an A haplotype chromosome. All 20 of the chromosomes carrying the –13.9kb*T allele also carried the –22kb*A allele (see table 3.3), although not all of the 28 A Haplotype chromosomes observed carried the –13.9kb*T.

Core Lactase haplotype	G-22kbA	C-13.9kbT	No of Chromosomes
A	A	T	20
A	A	C	2
A	G	C	6
B	G	C	13
C	G	C	3
D	G	C	2
E	G	C	1
F	G	C	1

Table 3.3 Derived alleles observed in a series of CEPH Northern French individuals

This result was consistent with expectations from previous studies, which had shown an association between the A haplotype and lactase persistence in Europeans (Harvey et al 1998), and is reported in Poulter et al (2003).[16]

*3.3.2 Association between –22kb*A and –13.9kb*T and the A Haplotype outside of Europe*

To establish whether the recently described alleles were found on the background of an A Haplotype outside Northern Europe, further data sets were used. Eight populations were tested for the –22kb*A and –13.9kb*T: Roma, North Indian, South Indian, South African Bantu-speakers, San, Malay, Japanese and Chinese. These groups had previously been characterised for the core lactase haplotype markers described by Hollox (2001)[17]. 390 individuals were used in the study, and

[16] See appendix C1

[17] A detailed description of the markers defining the core haplotypes can be found in Appendix B1

33 haplotypes were resolved using the computer program PHASE . It was possible to resolve some haplotypes by visual inspection of the data, for example, in cases where the genotypes were homozygous except for one marker. Table 3.4 shows the haplotypes for all eight populations. Interestingly, the –13.9kb*T allele was not observed in either of the two African populations, or the Japanese sample, and only at very low frequency in the Chinese and Malaysian samples.

Haplotype			*No. of chromosomes in population*							
Core Lactase Haplotype	*-22kb G/A*	*-13.9kb C/T*	*Roma n=162*	*North Indian n = 128*	*South Indian n= 68*	*Bantu n=50*	*San N=30*	*Malay n=192*	*Japan n=80*	*Chinese N=70*
A	**A**	**T**	**16**	**24**	**9**	**0**	**0**	**1**	**0**	**1**
A	**A**	**C**	**6**	**1**	**2**	**0**	**0**	**0**	**0**	**0**
A	G	C	55	25	19	6	2	95	29	26
B	G	C	47	28	14	0	1	27	5	7
C	G	C	16	34	20	15	1	35	13	13
D	G	C	2	0	0	0	0	0	0	0
E	G	C	5	1	0	0	0	3	3	2
G	G	C	4	6	2	0	0	4	2	4
H	G	C	0	0	0	0	1	0	2	0
I	G	C	0	2	0	0	0	1	0	0
J	G	C	3	0	0	1	0	1	0	1
J	**A**	**T**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**
K	G	C	3	0	1	0	0	1	0	0
M	G	C	1	3	0	2	0	0	0	0
N	G	C	0	0	0	0	0	0	2	2
O	G	C	0	0	0	4	1	0	0	0
(v) O	G	C	0	0	0	0	3	0	0	0
P	G	C	0	1	0	7	3	0	0	0
Q	G	C	1	2	1	4	1	4	0	2
S	G	C	0	0	0	2	3	4	2	2
T	G	C	0	0	0	1	1	0	0	0
U	G	C	0	0	0	3	5	15	21	10
V	G	C	0	0	0	0	1	0	0	0
W	G	C	0	0	0	0	0	0	1	0
X	G	C	0	0	0	3	3	0	0	0
Y	G	C	0	0	0	0	3	0	0	0
Z	G	C	0	0	0	0	0	0	0	0
1	G	C	1	1	0	0	0	0	0	0
4	G	C	0	0	0	0	1	0	0	0
6	G	C	0	0	0	1	0	0	0	0
7	G	C	0	0	0	1	0	0	0	0
(v) X	G	C	0	0	0	0	0	0	0	0
(v)	G	C	1	0	0	0	0	0	0	0

Table 3.4 Core Lactase Haplotype data for G-22kbA and C-13.9kbT polymorphisms in a series of populations
Haplotypes that are derived for G-22kbA and C-13.9kbT are shown in bold.
n = the number of chromosomes v = a new variant of a given haplotype

The 780 chromosomes described above reflect a wide degree of haplotypic diversity, however, as table 3.4 shows, both –22kb*A and –13.9kb*T alleles were mainly seen on the background of the A haplotype, corresponding with the association observed in the CEPH families from Northern France, and, also as in the CEPHs, not all A Haplotype chromosomes carry the derived alleles. There was

however, one exception; one Malaysian individual with haplotype BJ carried the –22kb*A and –13.9kb*T alleles. The PHASE software showed an equal probability of either the B or the J haplotype carrying the derived alleles, but previous haplotype reconstructions suggest that the J haplotype is more closely related to the A Haplotype (Hollox 2001, Appendix B1), differing only in one InDel polymorphism, a deleted base pair [A_8-552/-559 A_9].

3.3.3 Small family groups from eight populations

The final confirmation of haplotypic background for the –22kb*A and –13.9kb*A alleles used a series of samples of families collected from different populations by The Centre for Genetic Anthropology (TCGA). Haplotype was established using family pedigrees as previously described in the methods section 2.5 (Nejati-Javaremi and Smith 1996). These families had not been characterised for any of the core lactase haplotypes, and so, given the observations from 3.3.1 and 3.3.2, namely that the –22kb*A and –13.9kb*T alleles appeared to occur on the background of A Haplotype chromosomes, the T5579C polymorphism, which defines the A haplotype was typed. A Haplotype chromosomes carry a C at the T5579C site and all non A Haplotypes carry a T, with the exception of certain recombinant chromosomes as in the case of the E and F Haplotypes (Harvey et al 1998, Hollox et al 2001). Outside of Eurasia, a greater degree of recombinations breaks down the relationship between the 5579*C allele and the A Haplotype; the 5579*C allele is also found on the background of the following other haplotypes, specifically I, J, V, X, a, d, h, m and n.

An assay was optimised for the T5579C polymorphism (see section 2.4 and gel image fig 3.6). 'Condensed' three SNP haplotypes were constructed using this locus and also the G-22kbA and C–13.9kbT loci (see table 3.5). These are described with the alleles in the following order: G-22kbA, C-13.9kbT, T5579C. Four haplotypes were observed in total, GCT, GCC, ATC and GTT.

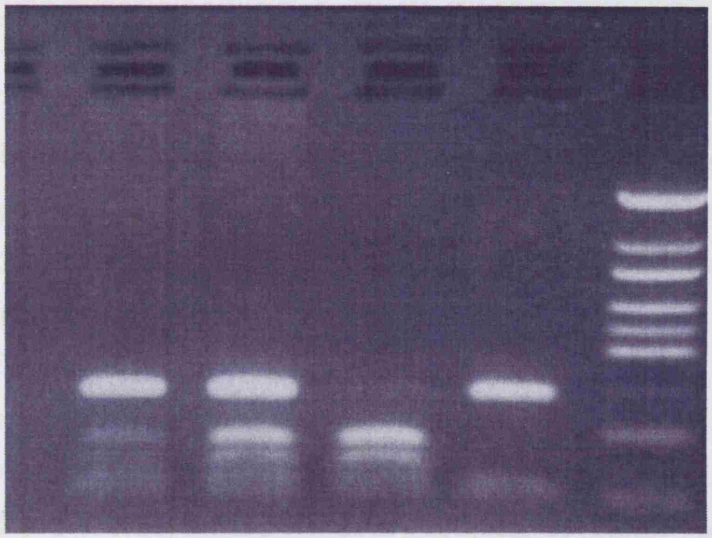


Fig 3.An example gel for the T5579C polymorphism
*The gel shows DNA derived from an enzyme assay digestion of a 143 bp PCR product. Lanes 1-4 show a series of samples, where the 5579*C gel phenotype band is 109bp, with a second product of 34bp, and 5579*T is 143bp. Lane 5 is a ladder of size markers*

Of the 541 chromosomes resolved, 157 carried the –13.9kb*T allele and all of these also carried the –22kb*A allele (see table 3.4 and fig 3.7).

Sample details		**Number of haplotypes observed**			
Population	**Number of chromosomes**	**-22kb *G -13.9kb *C 5579 *T**	**-22kb *G -13.9kb *C 5579 *C**	**-22kb *A -13.9kb *T 5579 *C**	**-22kb *G -13.9kb *T 5579 *T**
Irish	**65**	**1**	**2**	**62**	**0**
UK (English)	**64**	**13**	**4**	**47**	**0**
German	**60**	**19**	**8**	**33**	**0**
Jewish (Ashkenazi)	**96**	**63**	**25**	**8**	**0**
Armenian	**88**	**62**	**25**	**1**	**0**
Kuwaiti	**28**	**23**	**5**	**0**	**0**
Algerian	**21**	**7**	**7**	**6**	**1**
Ethiopian (Amharic)	**119**	**94**	**25**	**0**	**0**

Table 3.5 Condensed SNP Haplotypes comprising of G-22kbA, C-13.9kbT and T5579C SNPs observed in a series of families from different populations
The condensed SNP haplotypes and the MCM6 alleles are shown in red

In all cases, both the –22kb*A and –13.9kb*T alleles were shown to occur together on the background of the A haplotype, as defined by the presence of 5579kb*C allele, with one exception; a chromosome carrying the GTT (-22kb*G, -13.9kb*T, 5579*T) haplotype was observed in an Algerian family.

The 'condensed' SNP haplotypes, as shown in table 3.4, are comprised of three SNPs which have all been reported to associate with lactase persistence, (Harvey et al 1998, Hollox et al 2001, Enattah et al 2002). Frequencies of these varied between population groups as is shown in fig 3.7. There is a clinal pattern, with a higher frequency of the ATC haplotype in Northern Europe, which gets less further South and East. There also appears to be a discrepancy between the Ashkenazi Jewish population, and other European groups. Also, the Algerian population group shows, unusually, equal proportions of all three condensed haplotypes.

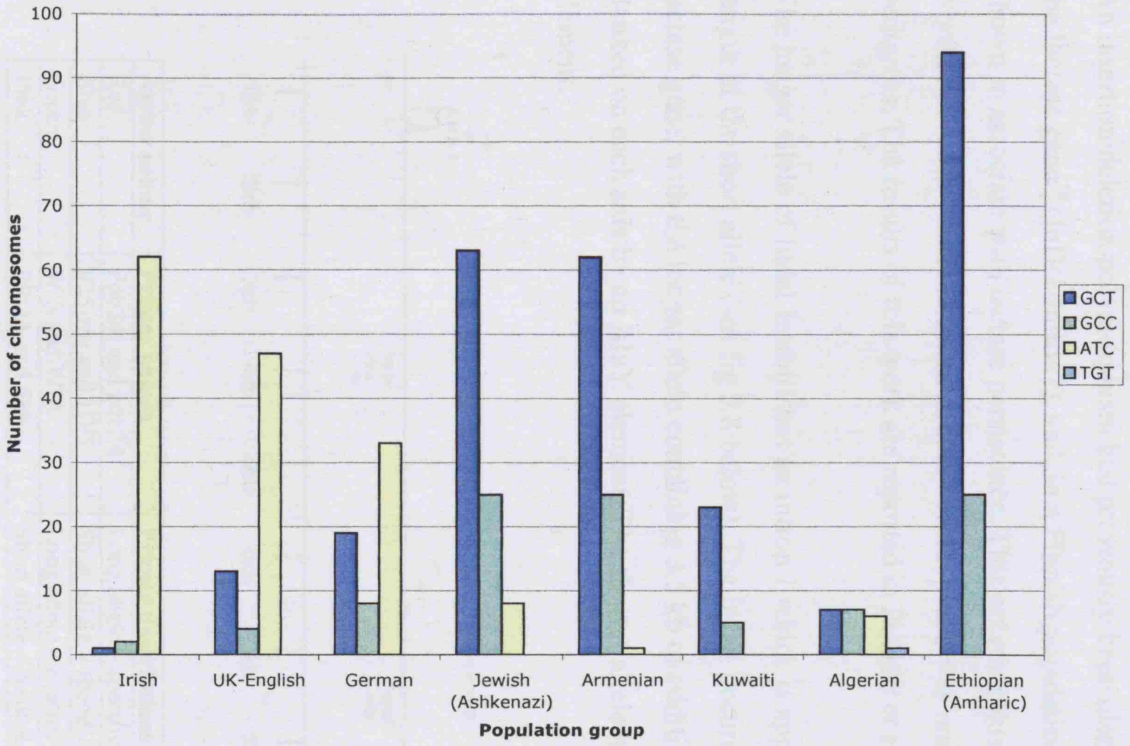
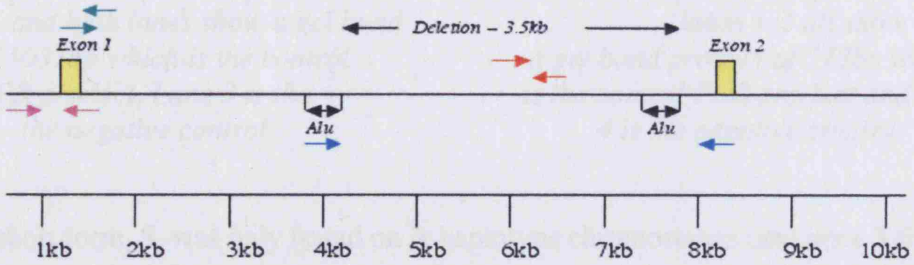


Fig 3.7. A bar graph to show the frequency of condensed SNP haplotypesin a series of populations, using three loci: G-22kbA, C-13.9kbT and T5579C, for the family populations. *The four observed haplotypes, GCT, GCC, ATC and GTT are derived as shown on table 3.4 and defined in section 3.3.3. The ancestral haplotype is assumed here to be GCT*

3.4 Examining a recently discovered Insertion/Deletion polymorphism on CEPH samples

An insertion/deletion polymorphism had previously been identified in intron 1 of the lactase gene[18] (InDel-Intron1), and, in a Finnish population, had previously been shown to associate with lactase persistence. This polymorphism[19] was typed in the Northern French CEPH families and, as before, phase was resolved using pedigrees. The results of this work are reported in Poulter et al (2003).

The longer allele of Indel-Intron1 has an intron 1 which is approximately twice the length of the short allele (see fig 3.8 below). The InDel occurs in intron 1 of the lactase gene, with the longer allele containing 3.5 kb of additional sequence, flanked on each side by an AluY element. The shorter allele has only one such Alu element.



Arrow colour	Primer names	Primer descriptions
Red	Pdb24r and pdb24	Long allele – specific primers
Blue	F25-rev and I1F3	Short allele – specific primers
Green	WIS and WIA	Long allele – control primers
Pink	PROA and 5FS	Short allele – control primers

Fig 3.8 The location of the InDel-Intron1 polymorphism and primers used to amplify the region

[18] The Insertion/Deletion polymorphism was first observed two years previously by Mark Poulter, a member of the Galton Laboratory

[19] NCBI ref: locus HUMLCT01, genebank identification no: M61834.

A PCR protocol was designed to identify the long allele, and each sample was also typed with a separate PCR protocol to identify the short allele (see section 2.4). In each case, as shown in figs 3.8 and 3.9, an internal set of primers was included to confirm the PCR was successful and to check the integrity of the DNA.

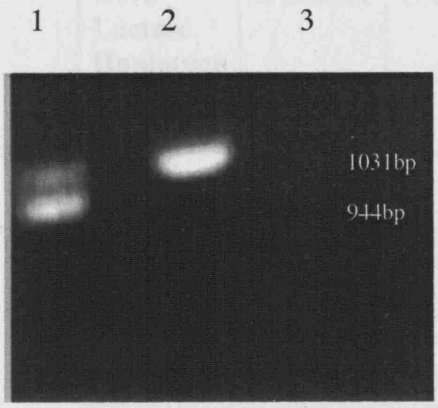


Fig 3.8 Gel image showing the gel phenotype for the short allele
Lane 1 shows a band of 944bp which is the short allele product and both lanes show a gel band of 1031bp which is the control PCR product. Lane 3 is the the negative control

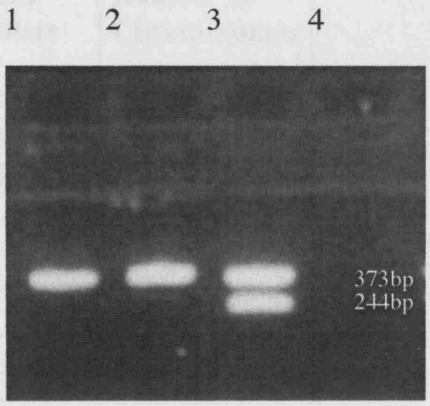


Fig 3.9 Gel image showing the gel phenotype for the long allele.
Lane 3 shows a gel band of 244bp which is the long allele product, and lanes 1-3 all show a gel band product of 373bp which is the control PCR product and lane 4 is the negative control

The short form, S, was only found on A haplotype chromosomes (see table 3.6), whereas the long form, L, was present on all non-A haplotypes, and also on two of the twenty-six A haplotype chromosomes, found in the Northern French. The two 'L' alleles occurring on the A haplotype were associated with the –13.9kb*C and –22kb*G alleles, but both these alleles were also present on two A haplotypes which carried the short version of intron 1.

Core Lactase Haplotype	G-22kbA	C-13.9kbT	L/S allele	Number of Chromosomes
A	**A**	**T**	**S**	**20**
A	**A**	**C**	**S**	**2**
A	**G**	**C**	**S**	**2**
A	G	C	L	2
B	G	C	L	13
C	G	C	L	3
D	G	C	L	2
E	G	C	L	1
F	G	C	L	1

Table 3.6 Association of the Long and Short alleles of the InDel-intron1 polymorphism with core lactase haplotypes. *Haplotypes were determined by inspection of the pedigrees. The derived alleles are shown on the table in bold*

3.5 Discussion

The association between C-13.9kbT and G-22kbA genotypes and lactase persistence phenotype as determined by the 'Gold Standard' method in a Finnish group was consistent with the findings of Enattah et al (2002). However, the existence of some British individuals characterised by enzyme assay as lactase persistence are homozygous for the 13.9kb C allele suggested that the association reported by Enattah et al (2002) is not as tight as might be expected. This may be because the -13.9kb T allele is not a causative mutation or is not the only causative mutation in Europeans. Confirmation by repeated sequencing of the genotype for one individual, sample 182, previously interpreted as homozygous persistent suggested that the –13.9kb*T allele might not explain lactase persistence phenotype in all individuals. This raises the importance of further investigation to determine if the –13.9kb*T allele is truly causal of lactase persistence in all individuals, or if, in

some populations, it is very tightly associated with a true and as yet unknown, causal mutation.

It seems likely that the –22kb*A and –13.9kb*T alleles are the derived forms of the polymorphism, since they associate with lactase persistence, which is generally accepted to be the derived form of the phenotype. Evidence to support this came from the great apes, all of whom were homozygote for –13.9kb*C allele.

This chapter investigated the haplotypic background of both derived *MCM6* alleles to place them in the context of what is already known about the evolution of polymorphisms in and around the lactase gene. The results from three groups of data strongly suggest that both –22kb*A and –13.9kb*T are found on the background of the A Haplotype, in Europe (France, England, Ireland, Germany and Ashkenazi Jewish), Asia (Roma, North and South Indian) and also Northern Africa (Algeria).

The –13.9kb*T allele was found at extremely low frequency in China (one case only, again showing association with the A Haplotype) and in two Malaysian individuals. In one of these cases, the allele associated with the A Haplotype, in the other, the core lactase diplotype was BJ, so the –13.9kb*T allele could have associated with either chromosome. The C-13.9kbT and G-22kbA assays were repeated for this sample, with the same results. It is probable in this case that the derived *MCM6* alleles associate with the J haplotype, since this is more closely related to the A (Hollox 2000).

It was not possible to determine haplotypic background for the alleles in sub-Saharan African, Kuwait or in Japan, as the –13.9kb*T allele was not observed in samples. In Africa, for both South African Bantu-speaking and San samples, previous lactose tolerance studies suggest that the range of lactase persistence frequency is between 0 and 5% (Cook and Kajubi 1966; Cook et al 1967; Cox and Elliott 1974; Nurse and Jenkins 1974; O'Keefe and Adam 1983; Segal et al 1983;

O'Keefe et al 1984). There are no lactase persistence studies specifically on the Amharic populations of Ethiopia. If the South African Bantu-speaking and San samples did not include any lactase persistent individuals, which is quite possible, then it would be expected not to see either allele in these two groups. Similarly, this may be the case in the Far Eastern populations, where the –22kb*A and –13.9kb*T were at very low frequency or completely absent. A study on phenotype in China suggests 80-100% of individuals are non-persistent (Yongfa et al 1984).

The haplotypic background of the InDel polymorphism in intron 1 was also characterised, as reported in Poulter et al (2003), with the shorter allele shown to occur on the background of the A haplotype. Given the frequency of the short allele, which is also associated with the –22kb*A and –13.9kb*T alleles, it seems likely that it is the oldest of the three polymorphisms described here which sub-divide the A Haplotype. It is probable also that the –22kb*A allele arose on the background of an A Haplotype carrying the short allele in intron 1, and, subsequently, the –13.9kb*T allele arose. If the –13.9kb*T allele is only linked to lactase persistence but not causative of it, it might be hypothesised, by looking at the pattern of association, that a causal mutation could have occurred after the –22kb*A allele but prior to the –13.9kb*T allele, although historic recombination events might possibly confound such an interpretation.

Chapter Four

The evolution of the lactase persistence phenotype in Africa

4.1 Introduction

The discovery of two SNPs closely associated with the lactase persistence phenotype in Europe (Enattah et al 2002), and the claim than one of them is causative of the lactase persistence phenotype has, potentially, two major repercussions. First, if the –13.9kb*T allele is the universal cause of lactase persistence humans then there is a strong diagnostic benefit. Establishing lactase persistence status is currently either invasive, as in the case of biopsy, or error prone and intrusive, as in the case of the indirect tests (for example, Kurt et al 2003). A method based on genotyping could potentially be more accurate, cost-effective and less intrusive to the patient. Secondly, the two polymorphisms could, from a population genetics perspective, help to explain the evolution of the lactase persistence trait and its global distribution.

Although the association between C-13.9kb*T allele and lactase persistence phenotype has not been conclusively shown in populations outside of Europe, the genotyping procedure is being used (at the time of writing) by the Medix laboratory in Finland, as a diagnostic test. A similar test is being proposed as a means of diagnosing the development of lactase non-persistence in children (Rasinpera et al 2004). Further studies have considered the C-13.9kbT polymorphism as a medical tool to predict risk factors: for example, one study provided evidence that –13.9kb*C/C genotype was associated with an increased risk of colorectal cancer in a Finnish sample (Raspinera et al 2005). The –13.9kb*C/C genotype has also been proposed to represent a genetic risk factor for bone fractures in old age in a Finnish sample (Enattah et al 2005), but not for stress fractures or bone density in a series of young male army recruits, also from Finland (Enattah et al 2005), and nor as a correlate for predicting risk factor for both diabetes types I and type II (Enattah et al 2004).

So far, the diagnostic use of such tests and scope of clinical associations using the C-13.9kbT polymorphism has been limited to Finland where the association was first observed. If an association can be demonstrated in outside of Europe, this might suggest a much wider application for genotyping procedures and inference.

In Africa, the distribution of lactase persistence is far more heterogenous than in Europe; although there is a general North – South cline, with lactase persistence frequency being higher in the Mahgreb, isolated groups of nomadic pastoralists with high frequencies of lactase persistence are frequent, and disrupt this pattern (for review, see Swallow et al 2003). Neighbouring population groups sometimes have different persistence frequencies, such as the Nuer and the Wolof (Arnold et al 1980) Some groups, particularly in the Nilo-Saharan language speaking regions, display only a low to moderate frequency of lactase persistence, despite a long history of pastoralism and milk drinking (Bayoumi et al 1981, 1983). In East Africa, along the Nile, there is a high frequency of lactase persistence amidst the Afro-Asiatic groups living parallel to the Nuba mountains, and this decreases both North and South with distance from this centre (Bayoumi et al 1981, 1983). A similar pattern can be observed in central Egypt, where lactase persistence frequency decreases North and South from the centre (Hussein et al 1982).

In their 2002 paper, Enattah and colleagues show that, in Americans of African and European ancestry, genotype frequencies for the C–13.9kbT polymorphism are consistent with previously reported levels of lactase persistence for the respective continents, and infer from this that there may be a wider diagnostic application (Enattah et al 2002). However, there is a possibility the observed association may be due to inappropriate comparison between African-Americans and continental Africans. There is also some evidence that the African-American population have a significant non-African component to their ancestry due to subsequent admixture (Parra et al 1998).

A further Finnish study tested, amongst others, 65 children with an African origin, three of whom were CT heterozygotes. Intestinal biopsy samples were available, and the authors observed a good correlation between phenotype established from these and the C-13.9kbT polymorphism even in children (Raspinera et al 2004). The authors suggested that the correlation between the non-persistent phenotype and –13.9kb*C homozygote genotype shows that the diagnostic potential of the polymorphism can be extended to Africa, and thus has a global relevance. However, this argument is only valid if the –13.9kb*T allele can similarly be demonstrated to associate with the lactase persistence phenotype in sample groups from Africa.

The findings described in chapter 3 suggest that the frequency of the –13.9kb*T allele in European, Roma and Indian populations might be consistent with reported frequency of lactase persistence, and the absence of the allele in the South African Bantu-speakers, and the San could perhaps be explained by low levels of lactase persistence in these populations. To investigate the association between lactase persistence and the –13.9kb*T allele further, a series of populations throughout the world[20] with varying frequencies of lactase persistence were typed for the C–13.9kbT polymorphism. The genotype data was then compared to previously reported frequencies. This chapter specifically investigates whether the –13.9kb*T allele can explain lactase persistence frequencies in sub-Saharan Africa, where some populations there do show intermediate to high frequencies of lactase persistence comparable with those observed in Europe[21].

If, as has been suggested, a strong selective pressure was involved in raising the frequencies of lactase persistence in populations with a history of pastoralism, a key question is whether or not this can be demonstrated by investigating frequencies of alleles associating with lactase persistence. The recently described *MCM6*

[20] Including those described in chapter 3

[21] For example, lactase persistence occurs in the Ga'ali at a frequency of 0.531, (Bayoumi et al 1981) compared with reported frequencies of lactase persistence in Greece, of 0.553 (Kanaghinis et al 1974)

polymorphisms are candidate alleles for this investigation. In this chapter the distribution of the –13.9kb*T allele in key sub-Saharan Africa populations is estimated. Some of the work shown in this chapter is reported in Mulcare et al (2004).[22] In addition to the published data, this chapter also includes data for the distribution of –22kb*A and 5579*T alleles in populations where lactase persistence frequency was known, and C-13.9kbT data for additional Ethiopian populations, a Berber group and a series of populations in the Extreme North Region of Cameroon.

4.2 Methods

4.2.1 Samples and anthropological information

Samples from a wide range of locations in sub-Saharan and one European control population were available from TCGA. A total of 1831 samples were initially typed for the C-13.9 kbT polymorphism and were classified into 34 distinct groups according to the country of collection and their self-declared cultural identity if this concurred with that ascribed to both their parents. Where there were less than 10 individuals of the same self-declared cultural identity they were classed as 'Other', and placed in a miscellaneous category for the region of collection. Individuals with partially unknown or diverse ancestry at the parental level were also classed as 'Other'.

The Ethnographic Atlas (Murdock 1967) was used to classify dependence on animal husbandry, milking practice, principal type of animals kept and nomadic status. A descriptive table of African pastoralists (Blench 1999) was used to classify pastoralist status according to a strict set of criteria: pastoralist groups were defined as those with a history of dependence on and migration with their animals. Information about first and second languages was available for each donor, and the Ethnologue[23] was used to investigate possible linguistic relationships between

[22] See appendix C2

[23] The Ethnologue can be found at: www.enthnologue.com

population groups. Ethnologue was also used to identify synonyms and different spellings for the same ethnic groups. Linguistic data and further ethnic classifications (such as birth place and home town) were also available for the maternal and paternal grandparents[24] of each sample donor and were referred to as necessary. For most groups, it was possible to clarify and confirm anthropological information and queries about particular criteria through discussion with the original sample collectors. This was especially important where published ethnographical studies were not available for populations, and/or where individuals in an 'Other' group practised similar farming economies. The anthropological background data relevant to lactase persistence is summarised in a table 4.1.

[24] Only data for maternal grandmother and paternal grandfather was available since the samples were originally collected for MTDNA and Y-Chromosome studies

Cultural Identity	Country	Nomadic Pastoralist status- (Blench 1999)	Animal husbandry Dependence and Milking (yes/no) (Murdock 1967)		Animals farmed	Nomadic status	Language	Language Phylum
Fulani (sedentary)	Cameroon	Yes	46-55%	Yes	Bovine	Settled	Fulfulde	Niger-Congo
Hausa	Cameroon	No	16-35%	No	Bovine	Settled	Hausa	Niger-Congo
Kwanja	Cameroon	No	-	-	-	-	Kwanja	Niger-Congo
Mambila	Cameroon	No	16-25%	No	Sheep/ Goats	Settled	Mambila	Niger-Congo
Nso (Nsaw)	Cameroon	No	6-15%	No	Sheep/ Goats	Settled	Nso	Niger-Congo
Yamba	Cameroon	No	-	-	-	-	Yamba	Niger-Congo
Other	Cameroon	-	-	-	-	-	-	-
Ibibio	Nigeria	No	6-15%	No	Bovine	Settled	Ibibio	Niger-Congo
Oron	Nigeria	No	(d)	-	-	-	Oron	Niger-Congo
Other	Nigeria	No	-	-	-	-	-	-
Chewa	Malawi	No	6-15%	Yes	Bovine	Settled	Chichewa	Niger-Congo
Ngoni	Malawi	No	6-15%	Yes	Bovine	-	Ngoni	Niger-Congo
Tumbuka	Malawi	No	6-15%	No	Sheep/ Goats	-	Tumbuku	Niger-Congo
Yao	Malawi	No	6-15%	No	Pets only	-	Chiyao	Niger-Congo
Other	Malawi	No	-	-	-	-	Bantoid language	Niger-Congo
Wolof (a)	Senegal	No	26-35%	Yes	Bovine	Settled	Wolof	Niger-Congo
Manjak	Senegal	No	-	-	-	-	Manj	Niger-Congo
Other	Senegal	-	-	-	-	-	-	-
Ga'ali	Sudan (North)	No	-	-	-	-	Arabic (Ja'ali dialect)	Afro-Asiatic

Cultural Identity	Country	Nomadic Pastoralist status- (Blench 1999)	Animal husbandry Dependence and Milking (yes/no) (Murdock 1967)		Animals farmed	Nomadic status	Language	Language Phylum
Shaigi	Sudan (North)	No	-	-	-	-	Arabic	Afro-Asiatic
Other (b)	Sudan (North)	Mixed	-	-	-	-	-	-
Dinka	Sudan (South)	Yes	46-55%	Yes	Bovine	Semi-nomadic	Dinka	Nilo-Saharan
Nuer	Sudan (South)	Yes	46-55%	Yes	Bovine	Semi-nomadic	Nuer	Nilo-Saharan
Other (b)	Sudan (South)	Mixed	-	-	-	-	-	-
Nuer	Ethiopia	Yes	46-55%	Yes	Bovine	Semi-nomadic	Nuer	Nilo-Saharan
Anuak (Anywak)	Ethiopia	Yes (c)	6-15%	Yes	Sheep/ Goats	Settled	Anuakigna	Nilo-Saharan
Amharic	Ethiopia	-	26-35%	Yes	Bovine	Settled	Amharic	Afro-Asiatic
Oromo	Ethiopia	-	-	-	-	-	Oromigna	Afro-Asiatic
A'ari	Ethiopia	-	-	-	-	-	A'Arigna	Afro-Asiatic
Other	Ethiopia	-	-	-	-	-	-	-
Mussesse	Uganda	No	-	-	-	-	Luganda	Niger-Congo
Other	Uganda	No	-	-	-	-	Bantoid language	Niger-Congo
N.European*	Ireland	-	36-45%	Yes	Bovine	Settled	English	Indo-European

Table 4.1 – Summary of anthropological and linguistic information for the thirty-four groups under investigation
Key for Table 4.1: 'a' includes 23 'Lebou' individuals; Lebou is a dialect of Wolof. b These groups include 12 individuals from traditional milk drinking peoples of known high frequency for lactase persistence (Beja, Misseri, Gomoeia, Shilluk (Bayoumi et al. 1981; Bayoumi et al. 1982)).; c probably don't drink fresh milk (Tarekegn unpublished, d identical to Ibibio in this respect,
'-' Information not available.
** This population is included since it was used as a positive control in subsequent analysis table 4.2*

4.2.2 Genotyping strategy

All samples were typed for the –13.9kb*T allele using the PCR and allele-specific restriction enzyme digest technique as described in the methods (section 2.4). For population groups where the frequency of the lactase persistence phenotype was known, the G–22kbA polymorphism and the T5579C polymorphism defining the A haplotype defining were also typed (2.4).

*4.2.3 Comparison of published lactose tolerance data and –13.9kb*T frequency*

In chapter 3, the association between lactase persistence phenotype and C-13.9kbT genotype was shown, in a Finnish group, to be comparable with the very tight association first described by Enattah et al (2002). Discrepancies that were observed could be explained when an error rate for the lactose tolerance tests, based on previous studies, was taken into account (Newcomer et al 1975, Howell et al 1981, Arola et al 1988, Peuhkeuri 2000, Kurt et al 2003). This being the case, it seemed likely that there would be a good correlation between genotype and phenotype in matched populations if this allele were really causative of lactase persistence in Africa.

An extensive literature survey was undertaken to collate studies reporting lactase persistence frequencies in sub-Saharan Africa. A series of review chapters were used initially to identify relevant papers (Holden and Mace 1997, Swallow and Hollox 2001, Swallow 2003), which were then included in a database if they fulfilled certain criteria. Studies using children under the age of 5, involving those with gastrointestinal problems (which might cause secondary hypolactasia) or where there was little or no classification as to ethnicity were excluded.

A statistical procedure, *GenoPheno* v.1.00, was designed by Dr. Mike Weale[25], (see 2.6.4) to take four possible sources of error into account: sampling error in the groups collected for the C-13.9kbT genotyping; sampling error in the published

[25] The procedure involved using a script written in the 'R' statistical programming environment, a copy of which is in the appendix A1 and also at www.ucl.ac.uk/tcga/software/

data on phenotype; error of false positive phenotype typing; error of false negative phenotype typing. The rate and direction of phenotyping error can be adjusted according to which of the indirect tests have been used (Newcomer et al 1975, Howell et al 1981, Arola et al 1988, Peuhkeuri 2000, Kurt et al 2003). Data from a series of unrelated Irish individuals (Fielding et al 1981) were compared, as a control, with C-13.9kbT frequency data for the TCGA genotyped Irish population (n=47).

4.3 Results

*4.3.1 Distribution of the –13.9kb*T allele in Africa*

The first and most noteworthy observation was that the frequency of -13.9kb*T was low or zero in most of the African groups tested. Only 5 out of the 33 sub-Saharan African groups showed any evidence of the –13.9kb*T allele, and in total it was found in only 26 out of 1784[26] sub-Saharan Africans, all but one of whom was living close to the same market town, Mayo Darle, Cameroon (see table 4.2). Ten of these 25 individuals were Fulani, four were Hausa, one was Mambila, and ten were of mixed ancestry at the parental level. The individual who did not come from the Mayo Darle area was from Senegal, and was also of mixed ancestry. Thus even in the groups where it is observed, the frequency of the –13.9kb*T allele is low compared to its frequency in the Irish control group, and in the Eurasian populations described in Chapter 3.

The highest –13.9kb*T allele frequency found in any African population was 0.139, in the Hausa samples from Mayo Darle (n = 18), compared to the higher frequency, 0.872, found in the Irish control group (n=47). It is also interesting that –13.9kb*T was not found in any of the East African samples examined, even though the data sets included many known pastoralists and groups with a high frequency of lactase persistence (see tables 4.1, 4.2, 4.3 and B1).

[26] This total does not include the Irish controls, and is a higher number than the reported data in Mulcare et al 2004 since additional Ethiopian populations were typed subsequently.

Country	Cultural Identity	Number of Samples	No. of CC	No. of CT	No. of TT	Frequency of the –13.9kb*T allele
Cameroon	Fulani (sedentary)	49	39	9	1	**0.112**
Cameroon	Hausa	18	14	3	1	**0.139**
Cameroon	Kwanja	70	70	0	0	0.000
Cameroon	Mambila	122	121	1	0	**0.004**
Cameroon	Nso (Nsaw)	126	126	0	0	0.000
Cameroon	Yamba	21	21	0	0	0.000
Cameroon	Mixed	128	118	9	1	**0.043**
Nigeria	Ibibio	110	110	0	0	0.000
Nigeria	Oron	44	44	0	0	0.000
Nigeria	Mixed	22	22	0	0	0.000
Malawi	Chewa	84	84	0	0	0.000
Malawi	Ngoni	14	14	0	0	0.000
Malawi	Tumbuka	58	58	0	0	0.000
Malawi	Yao	49	49	0	0	0.000
Malawi	Mixed (all Bantu)	58	58	0	0	0.000
Senegal	Wolof	69	69	0	0	0.000
Senegal	Manjak	93	93	0	0	0.000
Senegal	Mixed	19	18	1	0	**0.026**
Sudan (North)	Ga'ali	30	30	0	0	0.000
Sudan (North)	Shaigi	11	11	0	0	0.000
Sudan (North)	Mixed	88	88	0	0	0.000
Sudan (South)	Dinka	34	34	0	0	0.000
Sudan (South)	Nuer	13	13	0	0	0.000
Sudan (South)	Mixed	73	73	0	0	0.000
Ethiopia	Nuer	119	119	0	0	0.000
Ethiopia	Anuak (Anywak)	108	108	0	0	0.000
Ethiopia	Oromo	14	14	0	0	0.000
Ethiopia	Amharic	59	59	0	0	0.000
Ethiopia	A'ari	14	14	0	0	0.000
Ethiopia	Mixed	27	27	0	0	0.000
Uganda	Mussesse	22	22	0	0	0.000
Uganda	Mixed – all Bantu	18	18	0	0	0.000
Ireland	N. European	47	1	10	36	**0.872**
Total	All populations	1831	1759	33	39	

Table 4.2 Genotype and Frequency data for the C-13.9kbT polymorphism in the thirty-four population groups studied.
Populations where the T allele was observed are shown in bold

4.3.2 Does the distribution of the -13.9kb T allele predict lactase persistence frequency in sub-Saharan African populations?

In some cases, it was possible to find closely matching populations in the literature for which phenotype data was available (Table 4.3). This enabled a comparison between previously reported frequencies and predicted frequencies of lactase persistence based on the –13.9kb*T allele frequency. As discussed above, a statistical procedure was used to compare reported frequencies of lactase persistence with the frequency of the putatively causal –13.9kb*T allele in the comparable ethnic groups. Populations were matched to fulfil the following criteria: (1) same declared cultural identity and (2) residency in the same country or neighbouring state. Where neighbouring states were used, it was because the groups under investigation are nomadic and so were matched by region over which they migrate, often across country borders. Table 5.3 shows the results of the statistical comparisons[27]

[27] References for previously published data on lactase persistence are listed in the legend.

Group	Country of Genotyped Sample	No. of samples	Exp. F (Dig)	Country of Phenotyped Sample	No. of samples	Test method	Obs. F (Dig.)	Ref	P-value
Fulani (sedentary)	Cameroon	49	0.265	Nigeria	24	G	0.292	1	1
Hausa	Cameroon	18	0.305	Nigeria	17	G	0.235	1	0.749
Wolof	Senegal	69	0.086	Senegal	53	G	0.509	2	0
Ga'ali (Ja'ali)	Sudan (North)	30	0.068	Sudan	113	H	0.531	3	0
Shaigi	Sudan (North)	11	0.068	Sudan	42	H	0.381	3	0.025
Nuer	Sudan (South), Ethiopia	132	0.068	Sudan	23	H	0.217	4	0.030
Dinka	Sudan (South)	34	0.068	Sudan	208	H	0.255	4	0.001
European	Ireland	47	0.918	Ireland	50	G	0.900	5	1

Table 4.3: Comparisons with published lactose digester frequencies in matching populations, taking into account sampling and phenotyping error

Key for Table 4.3: G = blood glucose phenotyping test, H = breath hydrogen phenotyping test. Exp. F (Dig) = expected frequency of lactose digesters taking into account the test error rate by the method used in the matched population Obs. F (Dig) = observed frequency of lactose digesters in phenotyped sample Ref = reference for phenotyped samples 1, (Kretchmer et al. 1971), 2 (Arnold et al. 1980), 3 (Bayoumi et al. 1981), 4 (Bayoumi et al. 1982). 5 Blood glucose results only taken from (Fielding et al 1981), using a rise of >20mg/dl to define lactose digest. P-value = significance as determined by the statistical procedure described in 2.6 and 4.2

The predicted frequencies of lactase persistence, deduced from the frequency of -13.9kb*TT and -13.9kb*CT genotypes were significantly different from the reported frequencies obtained from published lactose tolerance data in all the sub-Saharan African populations with the exception of the Fulani and the Hausa (see table 4.3). Thus only in these two Cameroonian groups was the -13.9kb*T allele found at sufficient frequencies to explain the previously observed levels of lactase persistence. The Irish population, also included on table 4.3 as a control, shows no significant difference between the predicted and the observed frequencies of lactase persistence in the genotyped and phenotyped groups.

4.4 Alleles associating with Lactase Persistence in sub-Saharan Africa

If the –13.9kb*T allele is not the true causative mutation, the 'true' causal nucleotide change may be highly associated with the T allele, and may occur on a progenitor A Haplotype background. To investigate this possibility, the G-22kbA loci described by Enattah and colleagues (2002) was studied, and also the T5579C, where the 5579*C allele defines the A Haplotype. Populations for further investigation were chosen if lactase persistence frequency data was available. The Anuak were also included since, as a non-milk drinking neighbouring group to the Nuer, they provided a potential comparator group for frequency of the A Haplotype. Table 4.4 shows the genotype data for the three polymorphisms, all of which showed frequencies that were not significantly departed from Hardy-Weinberg expectation.

Sample Details			Allele Frequencies		
Population	**Country**	**No. of samples**	**-22kb*A**	**-13.9kb*T**	**5597*C**
Fulani	Cameroon	49	0.117	0.117	0.383
Hausa	Cameroon	18	0.139	0.139	0.500
Wolof	Senegal	41	0.000	0.000	0.256
Anuak	Ethiopia	108	0.000	0.000	0.208
Nuer	Ethiopia	119	0.000	0.000	0.206
Nuer	Sudan	13	0.000	0.000	0.077
Ga'ali	Sudan	30	0.000	0.000	0.217
Shaigi	Sudan	11	0.000	0.000	0.545
Dinka	Sudan	34	0.000	0.000	0.191

Table 4.4 Frequency data for a series of alleles associating with lactase persistence in a series of populations

By visual inspection of the data, and also using the program PHASE to determine haplotype, it was determined that the –13.9kb*T allele occurs on the background of a chromosome carrying the –22kb*A and the 5579*C alleles, as described in Chapter 3.

4.4.1 Distribution of the A Haplotype in different African groups

In contrast to the low frequency of –22kb*A and –13.9kb*T alleles across Africa, the 5579*C SNP was found both at a higher frequency in general and at noticeably different frequencies in the various populations as reported in table 4.4. An attempt was made to determine whether these differences in frequency could be correlated with any anthropological criteria. Comparison of pastoralists with non-pastoralists showed significantly fewer putative A haplotype chromosomes (5579*C) in the pastoralists, and fewer in the groups traditionally speaking Nilo-Saharan languages (fig 4.5). The populations and data were grouped as shown on table 4.1.

The Anuak proved problematic as a group; although defined by Blench (1999) as a pastoralist group, they were described by Murdock (1967) and the sample collector as an agriculturalist group with limited dependency on animal husbandry. For this reason, they were omitted from statistical analysis in the pastoralist comparison. The ratio of 5579*C: T alleles in the Anuak, 45:171, did

not differ significantly from a neighbouring group of pastoralists with a high dependency on animal husbandry, the Nuer ($p = 0.76$, χ^2 test).

Pastoralists and Non-pastoralists (i)

	Non-Pastoralists *(Hausa, Wolof, Ga'ali, Shaigi)*	**Pastoralists** *(Fulani, Nuer, Dinka)*
5579*T	136	326
5579*C	64	100
χ^2 value[28] = 4.69, df = 1, p = 0.0303		

Groups determined by language phylum (ii)

	Language Phylum		
	Niger-Congo (Fulani, Hausa, Wolof)	**Nilo-Saharan** (Nuer, Dinka, Anuak)	**Afro-Asiatic** (Shaigi, Ga'ali)
5579*T	137	439	57
5579*C	75	109	25
$\chi^2 = 21.21, df = 2, p =< 0.0001$			

Table 4.5 Tables (i) and (ii) to show the association between T5579 C (A Haplotype marker) and two anthropological criteria, pastoralism and language phylum. *The tables show the grouping of the populations by anthropological criteria also shown on table 4.1, and the raw data numbers used to determine p-values using a χ^2 Test*

A significant association could be observed between a higher frequency of A haplotype and pastoralism ($p = 0.03$), and a very significant relationship between language phyla ($p = < 0.0001$). It appears that Niger-Congo speakers, and, to a lesser extent, Afro-Asiatic speakers have proportionately more 5579*Cs than Nilo-Saharan speaking groups.

[28] The χ^2 test was calculated using the online statistical calculator at http://faculty.vasrar.edu/lowry/vassarstats.html

*4.4.2 Association between the –13.9kb*T allele and Fulani ancestry*

The results of the statistical comparisons described in 4.3.2 suggested that –13.9kb*T allele cannot predict or explain lactase persistence frequencies throughout sub-Saharan Africa. Further investigation of the individuals who do carry the –13.9kb*T allele was carried out to determine what factors best explained the observed distribution of the –13.9kb*T allele.

In all but two cases (one individual of mixed ancestry from Senegal and one Hausa) the –13.9kb*T carrying individuals, or one or both of their parents, spoke Fulfulde, a Fulani language. A Fisher's Exact Test showed that this association between carrying the -13.9kb*T allele and speaking Fulfulde (by individual or one of their parents) was significant both in the Cameroonian sample as a whole (**p= < 0.001, n=534**) and in the non-Fulani Cameroonians (**p=0.015, n=485**). The one individual from Senegal carrying –13.9kb*T was of mixed Wolof and Toucouleur (Tukulor) ancestry. 'Toucouleur' is a contentious term, but it is used by some anthropologists to describe mixed 'colour' or ancestry, specifically with a Peuhl or Fulani inheritance (for example, Arnold et al 1980).

4.5 Evidence for a historic introgression in the Fulani

These data suggest that a Fulani ancestry, as evidenced through stated cultural identity and linguistic associations, is the most significant correlate of the –13.9kb*T allele in sub-Saharan Africa. This in turn raises the question of how an allele, virtually absent in other sub-Saharan African populations, was found in Africans with a Fulani ancestry.

In 2002, Cruciani and colleagues identified a Y chromosome haplogroup at high frequencies in Cameroon that is generally absent in sub-Saharan Africa. Phylogeographic arguments suggest that this haplogroup ('R1*', using the nomenclature of the Y Chromosome Consortium 2002) has a non-African origin. Cruciani and colleagues found R1* Y chromosome at an average frequency of 40% in several Northern Cameroonian groups, including one

Fulani group (Cruciani et al. 2002). The authors suggest that a supra-Saharan origin for this haplogroup.

Y chromosome data for the Fulani[29] samples from Mayo Darle corroborate the findings of Cruciani and colleagues (Cruciani et al. 2002) in finding high frequencies of a haplogroup that is generally absent from Sub-Saharan Africa. In the samples from Central Cameroon that were examined here, evidence was found for the same haplotype using a marker that appears phylogenetically equivalent to those used by Cruciani et al. (2002) to delineate R1*, marker 92R7 (Mathias et al (1994). This marker occurred with a comparatively high frequency of 19% in the Fulani sample.

[29] The Y-Chromosome haplotypes for the Fulani and comparator populations were kindly made available by Dr.Mark Thomas for this observation, and first reported in Mulcare et al 2004 (see Appendix C2). Since these Y-Chromosome results were originally typed for a different study, although the samples themselves are from the same groups, due to some amplification failures in some cases the numbers are smaller.

Cultural Identity	Country	Total number of samples	Frequency of 92R7 derived haplotypes	Cultural Identity	Country	Total number of samples	Frequency of 92R7 derived haplotypes
Aghem	Cameroon	100	0.000	Sidama	Ethiopia	124	0.000
Bafut	Cameroon	67	0.000	Tigrian	Ethiopia	66	0.000
Bamileke	Cameroon	21	0.000	Wolaiyta	Ethiopia	108	0.000
Bamun	Cameroon	153	0.000	Zeyisse	Ethiopia	66	0.000
Bornu	Cameroon	20	**0.550**	Asante	Ghana	72	0.000
Fulani	**Cameroon**	**42**	**0.190**	Brosa	Ghana	67	0.000
Kwandja	Cameroon	77	0.013	Bulsa	Ghana	89	0.000
Mambila	Cameroon	157	0.013	Dagaati	Ghana	75	0.000
Mandara	Cameroon	12	**0.667**	Fante	Ghana	58	0.034
Massa	**Cameroon**	**35**	**0.143**	Frafra	Ghana	77	0.000
Mbo	Cameroon	162	0.000	Gonja	Ghana	75	0.000
Mbororo	Cameroon	14	0.000	Kasena	Ghana	85	0.000
Mousgoum	**Cameroon**	**35**	**0.286**	Kusaasi	Ghana	75	0.000
Nso	Cameroon	149	0.013	Mampruli	Ghana	66	0.030
Sara	**Cameroon**	**14**	**0.357**	Nankana	Ghana	36	0.000
Tang-Wimbum	Cameroon	14	0.000	Sefwi	Ghana	48	0.000
Warr-Wimbum	Cameroon	23	0.000	Sisaala	Ghana	74	0.000
Wum	Cameroon	10	0.000	Wali	Ghana	74	0.000
Yamba	Cameroon	44	0.000	Chewa	Malawi	92	0.000
Mixed group	Cameroon	113	0.044	Lomwe	Malawi	19	0.000
A'ari	Ethiopia	111	0.000	Ngoni	Malawi	17	0.000
Agaw	Ethiopia	255	0.000	Trumbuku	Malawi	62	0.000
Amhara	Ethiopia	408	0.002	Yao	Malawi	56	0.000
Anuak	Ethiopia	93	0.000	Yimbere	Malawi	19	0.000
Bench	Ethiopia	123	0.000	Annang	Nigeria	117	0.009
Benna	Ethiopia	15	0.000	Efik	Nigeria	128	0.008
Burne	Ethiopia	23	0.000	Ekoi	Nigeria	188	0.016
Dasenech	Ethiopia	26	0.000	Ewe	Nigeria	86	0.023
Daworo	Ethiopia	12	0.000	Ibibio	Nigeria	527	0.002
Dime	Ethiopia	37	0.000	Igbo	Nigeria	102	0.000
Dirashe	Ethiopia	29	0.000	Oron	Nigeria	141	0.000
Gamo	Ethiopia	200	0.000	**Arab**	**North Cameroon**	**185**	**0.314**
Gedeo	Ethiopia	116	0.000	Kanuri	North Cameroon	57	**0.491**
Gelila	Ethiopia	17	0.000	Kotoko	North Cameroon	360	**0.469**
Genta	Ethiopia	46	0.000	Tikar	North Cameroon	112	0.009
Gidole	Ethiopia	11	0.000	Lebou	Senegal	28	0.000
Goffa	Ethiopia	99	0.000	Manjak	Senegal	95	0.000
Gurage	Ethiopia	26	0.000	Tembo	Senegal	19	0.000
Hamar	Ethiopia	22	0.000	other	sub-Saharan Africa	275	0.007
Kaffa	Ethiopia	120	0.000	Acceron	Sudan	26	0.000
Koira	Ethiopia	34	0.000	Dabaina	Sudan	20	0.050
Konso	Ethiopia	103	0.000	Dinka	Sudan	31	0.000
Malli	Ethiopia	117	0.000	Ga'ali	Sudan	41	0.024

Cultural Identity	Country	Total number of samples	Frequency of 92R7 derived haplotypes	Cultural Identity	Country	Total number of samples	Frequency of 92R7 derived haplotypes
Masgan Gurage	Ethiopia	56	0.000	Jawamaa	Sudan	26	0.038
Mejjenger	Ethiopia	106	0.000	Korongo	Sudan - Nuba Hills	61	0.000
Nuer	Ethiopia	100	0.000	Laggori	Sudan - Nuba Hills	48	0.021
Ochollo	Ethiopia	10	0.000	Moro	Sudan - Nuba Hills	32	0.000
Oromo	Ethiopia	150	0.000	**Shaigi**	**Sudan**	**11**	**0.273**
Shekecho	Ethiopia	117	0.000	Mixed group	South Sudan	72	0.014
Sheko	Ethiopia	113	0.000	Mussese	Uganda	26	0.000

Table 4.6 A summary of the Y-Chromosome data available for a series of sub-Saharan population groups. *Comparatively high frequencies, those over 0.1 are shown in bold.*

Data was also available on all the Fulani samples for six Y chromosome microsatellites (DYS19, DYS399, DYS390, DYS391, DYS392 and DYS393, Thomas et al., 1999). Sufficient Y chromosome microsatellite diversity was observed among the 92R7 derrived Y-chromosomes to indicate that a single, recent founder could not have brought this haplogroup to this part of Africa. Among the fifteen chromosomes carrying the 92R7 allele, ten different microsatellite haplotypes were observed. It is therefore possible that the back-migration event which led to the introduction of R1* into sub-Saharan Africa (Cruciani et al. 2002) also brought the –13.9kb*T allele.

4.6 A historic introgression in Northern Cameroon

The evidence of Cruciani and colleagues (2002) strongly suggested that a Eurasian introgression was responsible for introducing a derived Y-Chromosome haplotype, rare in sub-Saharan Africa, into Cameroon. Specifically, their study showed that the Lake Chad region in the Extreme Northern district of Cameroon had a higher frequency of these haplotypes. The Y-Chromosome data shown on table 4.6 and also from Mulcare et al 2004 supported this observation; higher frequencies of the derived 92R7 haplotype were found in Cameroon, and in population groups near Lake Chad.

To resolve the history of Northern Cameroon further, and to investigate whether the –13.9kb*T allele continued to associate with a Fulani ancestry in the Extreme North of the region, a further sample selection were typed for the -13.9kb T allele.

4.6.1 Population History of groups collected from the Extreme Northern Cameroon Region

A series of locations in the Extreme Northern Region of Cameroon were used as collection points. This area incorporates part of Lake Chad and borders Chad and Nigeria. Different ethnic groups were spread between the collection sites, and there is no strong correlation between individual ethnic groups and the sampling location. The region itself is of historical interest, since three major migration routes introduced groups from the Mahgreb (defined here as the Northern coastal region of Africa bordering the Mediterranean) deeper into Africa, circa. 900 AD. The first of these, a Berber group known as the Sanhaja, opened up a trade route from Wadi Draa to the banks of Senegal, where some settled permanently. An offshoot of this group, the Tuareg, settled in the central section of the Sahara. A second trade route from Western Algeria to middle Niger, and a third from Tripolitania to Lake Chad both passed through Tuareg land, and in the 10th century, the kingdoms of Songhai, Kanem and Mali were, according to oral tradition, formed by Berber tradesmen (McEvedy 1995).

The origins of the Fulani are the subject of debate, but are thought to be outside Cameroon; based on ethnic traditions and linguistics, an origin in the Futa Toro region of the Senegal river basin has been proposed (Newman 1995), and where the descendents of the early Berber immigrants are thought to have lived. They are believed to have reached the south side of Lake Chad, where the samples described here originate, in the 16th Century, and there met the Chua Arabs who had migrated, originally from Upper Egypt, in the opposite direction across the Sahel corridor (McEvedy 1995).

The group of Fulani, (n = 14) collected from this region, are nomadic, unlike those collected from Mayo Darle who were a settled group. They have a higher

dependency on animal husbandry, ranging from 70-100% (Murdoch, 1967), gaining much of their wealth from herding and trading in cattle (Fanso 1989). Nomadic Fulani are also reported to have a higher frequency of lactase persistence, at 78%[30] (Kretchmer et al 1971). The Fulani of this region, often referred to as Mbororo (as distinct from the settled Fulani, or 'Fulbe'), dominate the area and have a history of aggression towards some of their neighbours (Fanso 1989). Their comparative wealth combined with their conversion to Islam of the region through a Fulani-based *jihad*, has endowed the Mbororo with some prestige. 'Fulbeisation', or the process of becoming Fulani is therefore common, and involves conversion to Islam, acquisition of cattle and learning of Fulani language, customs and ritual (Burnham, 1996). This current trend in the Northern Region may make admixture proportions difficult to determine.

In general, the non-Fulani groups in the Extreme Northern Region of Cameroon, typically the Chua Arabs, Kanuri and Hausa[31] are mobile traders with strong kin-bonds, but are vulnerable to a recent rise in banditry and have limited resources (Burnham 1996). The Chua Arabs are currently in a sustained, low-level conflict with the neighbouring Kotoko. The Kotoko are well established in the Extreme Northern Region of Cameroon, and oral records suggest their chiefdoms emerged before, and lasted through, Cameroon's colonial past (Fanso 1989). The Kanembou and Kanuri have a long history of animal husbandry, and, although both groups are less dependent on animals than the Fulani, having a dependency estimate of 16-35%, they are thought to practice milking (Murdoch 1967). All of these pastoralist groups use cattle as the primary animal that they are dependent on (Murdoch 1967).

A Berber group (n = 77) was used as a Supra-Saharan comparison population, and, given the migratory history of the group; the oral traditions of the Fulani suggest that they may have originated as a Berber group (Newman 1989). The samples used in this chapter were collected in Ifrane and Fez from Morocco.

[30] This observation was made using a small sample size of only 9 nomadic Fulani individuals

[31] Regrettably it was not possible to collect Hausa samples from this region

4.6.2 Samples and Method

638 samples were typed for –13.9kb*T allele. These were categorised both by ethnic identity, as described in section 2.1, and, as before, a miscellaneous category was included for individuals with mixed ancestry at the parental level. The samples were originally collected from a series of ten sites in the Extreme Northern Region, and were also grouped by these locations. Samples were typed as before, (see section 2.4) and statistical analysis was carried out as described in section 2.7, appendix A1 and as before in this chapter.

4.6.3 Results

*4.6.3.1 Distribution of –13.9kb*T allele in the Extreme Northern Cameroon Region*

Table 4.7 summarises the genotypes and frequencies observed in the populations collected from Cameroon.

Population Group	Total number of samples	Frequency of –13.9kb*T allele
Chua Arab	158	0.035
Fulani	14	0.429
Kanembou	7	0.071
Kanuri	40	0.000
Kotoko	298	0.008
Mixed	121	0.017
Total in Extreme North Region of Cameroon	638	0.026
Berbers	77	0.136

4.7 A table to show the frequency of the –13.9kb*T allele in a series of populations from the Extreme Northern Region of Cameroon

As before, the linguistic data for sample donors, parents and grandparents were considered to identify possible cases of historic admixture and/or a Fulani ancestry, and/or, any other association with a particular group. In total, 27 individuals carried –13.9kb*T allele either as homozygotes or heterozygotes. Of these, 10/27 either spoke Fulani or had parents who did, and 16/27 either spoke Arabic or had parents who did. The remaining -13.9kb*T carrying individual was French-speaking, with French speaking parents.

The -13.9kb*T allele was found at highest frequency in the Fulani (0.429), approximately four times higher than the frequency observed in the settled Fulani of Mayo Darle, of 0.112. Using the statistical procedure implemented in *GenoPheno* to compare predicted and observed levels of lactase persistence in a matched group while taking into account the various sources of error (see section 4.2.3), no significant difference was found for the nomadic Fulani (p=

0.71958). Lactase persistence frequency data was taken from Kretchmer et al (1971). The Berber samples showed a frequency of –13.9kb*T allele that was intermediate between the two Fulani groups, settled and nomadic, perhaps lower than might be expected given that the group traditionally has a high dependence on animal husbandry, a long-term history of pastoralism and milk drinking (Murdoch 1967).

*4.6.3.2 Does Geographical Location best explain the distribution of –13.9kb*T allele?*

Cruciani and colleagues observed the haplotype R* in Y chromosomes from a range of populations, not just the Fulani, and, the data above suggests that the –13.9kb*T allele does occur in non-Fulani groups in this region (albeit at low frequency). It was possible that the distribution of –13.9kb*T alleles might mirror the distribution of Y Chromosome haplotypes observed by Cruciani and colleagues (2002), such that the data could be explained by geography. The table below, (table 4.8), shows the distribution of the C-13.9kbT genotypes as arranged by sampling location rather than ethnic identity. Sampling locations are distributed Northwest to Southeast.

Place of Collection	Longitude	Latitude	Number of samples	-13.9kb*T allele frequency
Afade	12.233	14.633	76	0.046
Blangafe	12.383	14.300	74	0.034
Goulfey	12.383	14.900	68	0.000
Hile Alifa	12.700	14.300	33	0.076
Kousseri	12.083	15.033	68	0.007
Logone-Birni	11.783	15.100	45	0.011
Makari	12.583	14.467	89	0.011
Maltam	12.183	14.817	59	0.025
Sadigo	12.317	14.517	66	0.023
Waza	11.400	14.567	60	0.050

Table 4.8 Summary of genotype and frequency data for the C-13.9kbT polymorphism across collection sites

There does not appear, from the distribution of allelic frequencies shown above, to be a geographic cline to the distribution of –13.9kb*T alleles in the regions samples. This can be seen more clearly on figure 4.1.

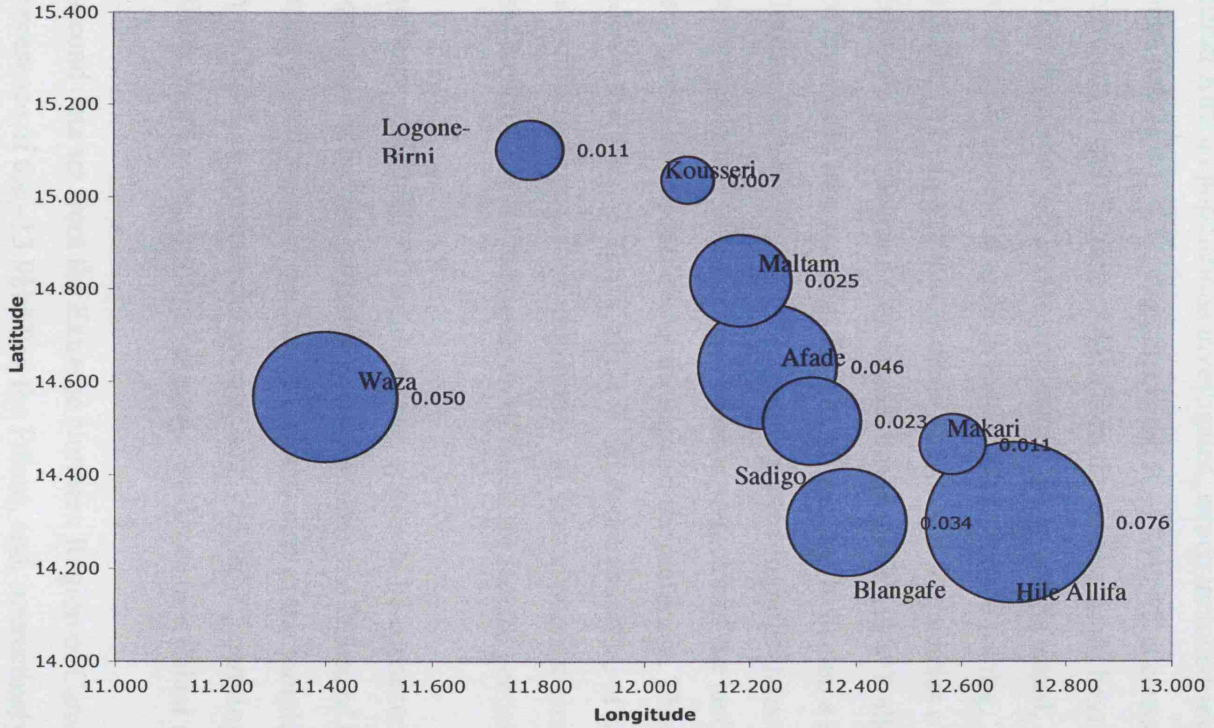


Fig. 4.1 A bubble-graph to show the frequency of the –13.9kb*T allele against latitude and longitude of collection site in the Extreme Northern Region of Cameroon. *The frequency value is shown by the comparative size of each bubble, and also shown as actual raw data by the side of each data point, where the collection location is also given. Goulfey is not shown, since no –13.9kb*T alleles were found there.*

4.7 Discussion

The low frequency, and complete absence of –13.9kb*T in most of the sub-Saharan African populations investigated, which included several milk-drinking groups (tables 4.1, 4.2), was surprising. If –13.9kb*T is causative of lactase persistence in Sub Saharan Africa then the results might have been explained by the low frequency of lactase persistence in the groups tested, and sampling error as well as experimental error in the lactase tolerance testing. To investigate this possibility, a comparison was made of reported frequencies of lactase persistence and observed frequencies based on –13.9kb*T allele which took such error rates into account. These comparisons indicated a highly significant difference between the observed and expected findings in most cases, indicating that the –13.9kb*T allele is not a reliable predictor of the lactase persistence phenotype in most populations throughout the continent. This is persuasive evidence that the –13.9kb*T allele is either not causative of lactase persistence, or is not the sole causative mutation in all human populations. Thus it is possible that there is heterogeneity of cause for lactase persistence.

Frequency of –13.9kb*T did, however, appear to be consistent with reported frequencies of lactase persistence in the Fulani and Hausa of Mayo Darle. Further investigation of the family backgrounds of the sample donors indicated some admixture between the groups, with –13.9kb*T carrying Hausa individuals showing, through language, evidence of a Fulani ancestry.

A second data set from the Extreme Northern Region of Cameroon confirmed the presence of the –13.9kb*T in the Fulani, and, interestingly, with the exception of the Kanuri, the allele is present at low frequencies in the neighbouring groups of the Extreme North Region. Although lactase persistence frequency data was not available for the other groups in the Extreme Northern Region, it is known that there is a long history of pastoralism and milking in the area, and, as before, the –13.9kb*T allele is at a lower frequency than might be expected in all other groups apart from the Fulani, assuming that it is causative

of lactase persistence. In the nomadic Fulani, as in the Fulani of Mayo Darle, the frequency of the T allele was not significantly different from the lactase persistence frequency matched set, suggesting, with the caveat of the small sample sizes in both the genotype data set (14) and phenotype data set (9), that it may account for lactase persistence in this population.

The Y-Chromosome data from Cruciani et al (1999), and that generated by the TCGA lab, both suggest an introgression event, probably from Eurasia, in the Extreme North Cameroonian region. Y chromosome haplogroup R1* is also found at high frequency in several non-Fulani groups in the Extreme North Province of Cameroon. However, the –13.9kb*T allele is not. This being the case, the presence of the –13.9kb*T at low frequency in non-Fulani groups here may have arisen due to local admixture between the Fulani and their neighbours[32]. The discrepancy between the distribution of the non-African Y-Chromosome haplogroup and the –13.9kb*T allele suggests that the demographic processes leading to the existence of the –13.9kb*T allele in Cameroon may be not be the same as those leading to the Y chromosome introgression, but could instead relate more specifically to Fulani migration history. Alternatively, the demographic process may be the same but subsequent drift has resulted in asymmetric distributions of these non-African genetic markers.

The presence of the –13.9kb*T allele at a similar frequency in the Berbers and the Mahgreb comparator group is interesting since the former population has a long history of trading and mixing with Eurasia (McEvedy 1980) and also the Fulani oral history claiming descent from a Berber sub-group (Newman 1989). The combined results from C-13.9kbT and the Y chromosome analysis suggest a complex demographic history for this part of Cameroon, which includes at least one major introduction of genes from outside the region.

[32] Personal communication with sample collectors suggests this scenario is possible

The data collected shows that the 5579*C allele exists at significantly higher frequencies in Africa than either of the two MCM6 derived alleles. It is noteworthy that in the Fulani and the Hausa the –13.9kb*T allele associates with the 5579*C allele and so may occur on the background of a putative A Haplotype here as well as in Europe. However, given a greater expected heterogeneity for lactase haplotypes in sub-Saharan populations (for example, Hollox et al 2002), it is likely that the 5579*C allele may associate with more haplotypes in the groups under investigation. Specifically, as discussed in chapter 3, the haplotypes E, F, I, J, V, X, a, d, h, m and n also carry 5579*C allele (Hollox et al 2001).

The 5579*C alleles showed a negative correlation with pastoralism, and a positive association with Niger-Congo language phylum. It is also noteworthy that the frequencies for the 5579*C allele in the milk drinking groups for which published lactose tolerance data is available (specifically, the Nuer, Dinka and Wolof) are lower than would have been expected if a causal nucleotide change was also associated with the A haplotype in these groups. These preliminary observations could point to a possible independent origin of lactase persistence in sub Saharan Africa.

Whether or not –13.9kb*T allele is causal, it seems probable that the C to T transition at –13.9kb occurred in a non sub-Saharan African population that contributed to the current population of Europe. If this is the case, then its presence in Cameroon, and especially in people of Fulani cultural identity or with Fulfulde speaking ancestry, can be explained by introgression from outside sub-Saharan Africa.

Chapter Five

The Evolution of Lactase Persistence in Eurasia .

5.1 Introduction

The –13.9kb*T allele has been shown to correlate well with lactase persistence in Northern Europeans (Enattah et al 2002, Poulter et al 2003, Kuokkenan et al 2004). However, this association is not observed throughout sub-Saharan Africa where it is almost completely absent with the exception of the Fulani, and those with Fulfulde (Fulani) speaking families (Mulcare et al 2004, Bersaglieri et al 2004). This chapter investigates how far the correlation between –13.9kb*T and lactase persistence holds throughout Eurasia, and also summarises the global distribution of the lactase persistent trait and alleles associating with lactase persistence. Characterising the origin and movement of the –22kb*A and –13.9kb*T alleles may, by proxy, provide information about the demographic history of the populations where the derived alleles have been observed at high frequencies, and the evolution of the lactase persistence phenotype itself.

The spread of agriculture from the Near East is well documented, but there is some disagreement amongst archaeologists as to whether the spread of a Neolithic cultural package across Europe also involved colonization of the continent by Neolithic farmers or involved the spread of the idea of farming without major demographic population movements. The first scenario, which assumes a significant migration of Neolithic populations into Europe, is sometimes described as the Demic Diffusion model (DDM), a term originally coined by Ammerman and Cavalli-Sforza (1984). The conflicting view, the Cultural Diffusion model (CDM) proposes the mass adoption of agricultural techniques by Paleolithic and Mesolithic communities occurred without substantial movement of peoples (for example, Richards et al 2000) Given that these two models conceive dramatically different contributions of early Neolithic farmers to the modern European gene pool, they have attracted the interest of population geneticists seeking to evaluate the likelihood of both theories.

In their seminal work, 'The History and Geography of Human Genes', Cavalli-Sforza, Menozzi and Piazza (1994) produced a map of Europe showing the first component (comprising approximately 26% of the variation in allele frequencies) of a principal component analysis. This map used the frequency distribution of 95 genetic loci in Europe groups to demonstrate a cline running North-West to South-East across Europe, which correlated well with traditional views on the trajectory of spread of Neolithic farmers from the Near East across Europe and which was used by the authors to support the Demic Diffusion model. The DDM has, partly as a result of such work, received wide support, and has influenced more popular views regarding the colonization of Europe, including the map taken from the Times Concise Atlas of World History (1994) shown in section 1.3.

More recently, other studies have similarly used molecular analysis to support the Demic Diffusion model (for example, Chikhi et al 1998). One study using Y-Chromosome markers took 3,616 chromosomes from European and surrounding populations, and investigated diversity using 11 biallelic loci. The data showed a highly non-random distribution of haplogroups, showing clinal variation in Europe consistent with a Near-Eastern immigration event, again correlating genes with geography (Rosser et al 2000). The authors also suggested that internal demographic events within Europe, specifically a distinct population movement from the Northern Black Sea region, revealed a complex history for the continent.

However, other research has shown a distribution of haplotypes more consistent with the Cultural Diffusion model, which predicts an older, Paleolithic base for the European gene pool. In particular, several mitochondrial DNA studies have not shown such a clinal distribution of allelic frequencies, and using a phylogeographic approach (for example, Richards et al 1996, Xiao et al 2004) have suggested that as much as 85% of European mtDNA lineages originate in the Upper Paleolithic, with only 13% likely to be from Middle Eastern Neolithic farmers (Richards et al 2000, Richards 2003).

One Y-Chromosome studies used twenty-two binary markers on the Y-Chromosome to show that >95% of 1007 chromosomes (of European and Middle-Eastern origin) could be accounted for by just ten lineages (Semino et al 2000). Two of these ten lineages were described as being 'present in Europe since Paleolithic times', whereas the remaining eight were though to enter Europe during later, independent migrations into the continent from the Middle East and the Urals (Semino et al 2000). This study suggested a comparatively limited contribution from Near Eastern farmers, lower than that which would be expected under the DDM, possibly as low as 22%.

Re-analysis of the same data set by Chikhi and colleagues (2002) suggested that Paleolithic populations contributed less than 30% to the European gene pool. Their starkly different conclusions rest upon a different methodology; they used admixture to distinguish between the relative contributions of Paleolithic and Neolithic gene pools, assuming that modern population groups from the Near East and Basque area could be used to represent descendants from the Neolithic and Paleolithic population groups respectively (Chikhi et al 2002). This suggests that different approaches to the analysis of genetic variation strongly influence the conclusions that can be drawn from a given data set.

Some studies suggest that the movement and changing identities of modern Europeans blur historic genetic associations (Pluciennik 1996), and, similarly that recent gene flow can confound historic demographic patterns (Ray et al 2003). Fix (1996) suggests that natural selection has affected modern gene frequencies in Europe – and, indeed, that the lactase persistence phenotype is a candidate for such selection.

A very recent study undertook a series of extensive demic computer simulations of the colonization of Europe by Neolithic farmers (Currat and Excoffier 2005). Their model takes into account various migration patterns, using different values for potential admixture and competition with the local Paleolithic populations. The authors test the probability that clinal patterns, traditionally seen as evidence for DD model, are equally probable under a model of cultural

diffusion since they could equally be caused by earlier colonizations of the continent in the Paleolithic. The authors also comment that many studies using SNP diversity were vulnerable to ascertainment bias, since only SNPs with high frequencies in Europe were used and it is these markers that are more likely to show clinal distributions (Currat and Excoffier 2005).

It is also possible to consider the movement of cattle grouped by geographic region as a proxy for the movement of their owners. A study used mitochondrial DNA from 392 domesticated cows from Europe, the Near Eastern and Africa to determine the origin of European domesticates (Troy et al 2001). The study compared haplotypes constructed from these extant animals with those from four extinct wild British oxen. Network analysis suggested that a Near Eastern origin was likely for the European cattle, which were distinctly different from the haplotypes of the extinct native oxen (Troy et al 2001). An earlier study similarly investigated gene frequencies in cattle, suggesting a specific route for the introduction of cattle into Europe from the Near East via that Balkans (Medjugorac et al 1994). As discussed in the introduction, Beja-pereira and colleagues studied the geographic distribution of variation in genes encoding milk proteins in European cattle breeds (2003). Their observations suggested a strong correlation between milk gene diversity and lactase persistence frequency in European groups, indicating a gene-culture coevolution between cattle and humans.

Given the complexity of European demographic history, the variation in and around the lactase gene can contribute to the observations made regarding the Neolithic cultural package that reached North-Western Europe approximately 4000 years ago

5.2 Methods and Samples

Genotyping techniques and statistical procedures were, as described before in methods 2.4 and chapter 4. As before, a statistical procedure taking into account four sources of sampling error was used to compare observed and expected frequencies of lactase persistence, where the expected frequency was calculated by assuming that the –13.9kb*T allele is causative of LP.

To create global interpolated distribution maps, data points were plotted against longitude and latitude co-ordinates. In the case of the TCGA and Galton samples, these co-ordinates were those of the sample collection points. For published data, collection points were plotted either as given in the literature, or, if no precise location was specified, the nearest city in the general area. If no information at all about collection site was given, co-ordinates for the ethnic group were taken from Murdoch et al (1967). A statistical program using Generic Mapping Tools (GMT) software was used to inter-polate the data points and to generate a gradient smoothed map[33].

For the global interpolated map of allelic frequencies, C-13.9kbT data generated from a total of 3243 samples from 94 populations was used. In addition to using 1607 samples described in chapters 3 and 4, a series of 467 unrelated individuals from 15 populations throughout Eurasia were available from TCGA. Published frequencies from 50 populations were also used (Bersaglieri et al 2004). For the global interpolated map of lactase persistence frequencies, the literature review undertaken for Africa (see 4.2) was extended to encompass all world populations. As before, studies that included children, low sample numbers or individuals with gastro-intestinal symptoms were excluded. Data from 160 distinct populations was used in total. Studies involving immigrant populations (such as Americans with Singaporean ancestry) were only used if detailed information and family history was provided, such that sample donors could be classified by their country of origin.

[33] This software is available from http://gmt.so-est.Hawaii.edu/

5.3 Results

5.3.1 Frequency distribution of alleles associating with lactase persistence

The frequency of the derived alleles is shown for G-22kbA, C-13.9kbT and T5579C, (table 5.1) for each of the populations typed in this chapter[34]. Appendices B2 and B3 show the complete lactase persistence frequency data and -13.9kb*T allele frequency data against longitude and latitude co-ordinates.

[34] The data from Bersaglieri et al 2004 is included for figs 5.1,5.3 and 5.4 to illustrate variation of frequency in Eurasia, but not in table 5.1 which shows only data generated during this project.

Sample details			Frequency of derived allele		
Population	**Country**	**Number of chromosomes**	**22kb*A**	**-13.9kb*T**	**5579*C**
Pashtu	Afghanistan	16	0.125	0.125	0.375
Tadjik	Afghanistan	98	0.112	0.102	0.429
Uzbek	Afghanistan	76	0.118	0.079	0.526
Anatolian Turkish	Turkey	98	0.082	0.031	0.429
Assyrian	Syria	80	0.050	0.038	0.313
Greek	Greece	82	0.159	0.134	0.488
Iranian	Iran	90	0.044	0.044	0.300
Israeli Arab	Israel	36	0.000	0.000	0.361
Israeli Bedouin	Israel	26	0.000	0.000	0.308
Jordanian Bedouin	Jordan	40	0.075	0.075	0.250
Palestinian Arab[a]	Palestine	34	0.029	0.029	0.25 (n=20)
Saudi Bedouin[a]	Saudi Arabia	86	0.000	0.000	0.106 (n=66)
Ukrainian	Ukraine	92	0.293	0.217	0.500
Azeri	Azerbaijan	44	0.023	0.023	0.182
Uzbekistani	Uzbekistan	36	0.000	0.000	0.306

Table 5.1 A table to show the frequency of derived alleles associated with lactase persistence
'a' indicates that for these two populations, it was not possible to type all the samples for the C5579T locus and so the actual sample number for this polymorphism is given with the frequency

The bar graph (fig 5.1) shows the frequencies of the –13.9kb*T allele as before, labelled according to ethnic group and orientated by longitude, west to east. As seen before on the maps, there is a general cline from the North-West to the South-East, although the frequencies observed in the Indian and Afghanistan populations tested were higher than some of those observed in the Mediterranean and Eastern European groups. The Bersaglieri data set (2004) for the Israeli Bedouin was used for the bar graph since the sample size was higher, though it should be noted that in the TCGA sample set the –13.9kb*T allele was absent. The allele was also absent in the Israeli Arab population. However, it was present in other parts of the Middle East, notably in the Iranian, Palestinian and Jordanian Bedouin groups.

The frequency of the –13.9kb*T allele was high in the Orcadians and a group of French Basques, both thought to be more related to the original Mesolithic communities in Europe (Wilson et al 2001). The –13.9kb*T allele was also found at very low frequency in some populations in China.

It is also noteworthy that, although frequencies of lactase persistent individuals diminish into South-East Europe compared to the North-West of the continent, they are elevated slightly in the Caucasus region of Eastern Europe, and also in Pakistan and Afghanistan.

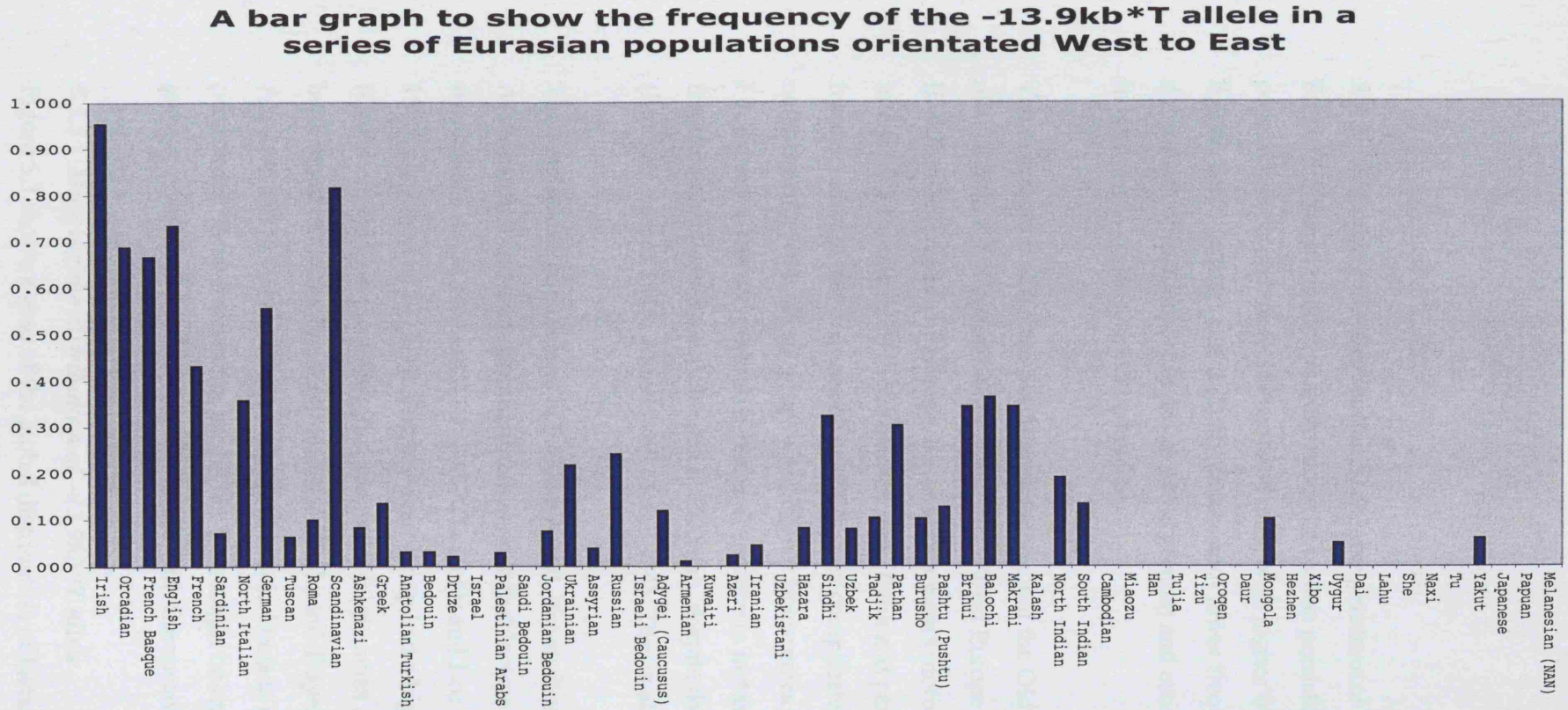


Fig 5.1 A bar graph to show the frequency of the –13.9kb*T allele in a series of Eurasian populations

The X axis of population groups is roughly arranged by longitude co-ordinate from West to East and the Y axis shows frequency

5.3.2 Mapping the distribution of lactase persistence in Eurasia

Figure 5.2 shows published frequencies of lactase persistence from a series of populations plotted on an interpolated map, with higher frequencies shown in a lighter shade getting progressively darker with lower frequencies. The map uses data points of frequency (shown as white glyphs) and estimates the distribution in between to create a smoothed gradient.

The distribution of lactase persistence throughout the Old World shows, as discussed in 1.2, a general cline from North-West Europe across to the South-East Europe. Outside of Europe, the pattern appears more complex; there are intermediate frequencies in the Middle East, India and parts of Africa. Interpretation of the frequencies on the map is complicated by the presence of neighbouring populations with very different frequencies of lactase persistence. For example, there appears to be large difference in lactase persistence frequency between the populations in Pakistan, despite their close proximity (Rab and Baseer 1976, Ahmad and Flatz 1984). This is also observed in Africa.

The map covers populations in the Old World, but studies on native North American and Canadian populations suggests that lactase persistence frequencies range between 0.19 – 0.37 (Leichter and Lee 1971, Bose and Welsh 1973, Caskey et al 1977). A Puerto-Rican group had a higher frequency of 0.46 (Goldman and Corcino 1976). In Greenland, frequencies of 0.12 –0.21 have been reported for the Inuit communities (Gudmand-Hoyer and Jarnum 1969, Asp et al 1975). Given that the ancestral trait in humans is thought to be non-persistence, these frequencies are higher than might be expected for these groups, possibly due to historic admixture with Europeans.

*5.3.3 Mapping the distribution of –13.9kb*T allele*

Figure 5.3 shows a map of the global distribution of lactase persistence as predicted by the –13.9kb*T allele using frequency data from table 5.1 and from Bersaglieri et al (2004), and figure 5.4 shows the raw frequency data for the –

13.9kb*T allele. To generate figure 5.3, Hardy-Weinberg equilibrium was assumed and an error rate associated with the indirect testing of lactase persistence phenotype derived from previous studies as described in 5.2 and 2.6.5. Comparing this map of lactase persistence frequencies predicted by the –13.9kb*T allele with figure 5.2, it seems clear that in Europe, there is a good correlation between the –13.9kb*T allele and lactase persistence. The correlation appears strong throughout central Europe, but appears to break down not only in Africa as previously discussed, but also in Greece, Turkey and the eastern region of Eurasia. From this data set, it seems probable that only in North-Western Europe is the relationship between lactase persistence and –13.9kb*T allele strong enough that the latter can be used to predict the former. Unlike the distribution in Africa, the –13.9kb*T and –22kb*A alleles are both present throughout many Eurasian populations, albeit at varying frequencies. Surprisingly, the –13.9kb*T allele was rare in the Bedouin groups tested, despite their long history of pastoralism (Cook and Al-Torki 1975). Data from Bersaglieri et al (2004) was consistent with this: they found a frequency of 0.031 for the allele in their sample set (n= 98).

The highest observed frequency of the –13.9kb*T allele in the dataset was is in Ireland (0.954). It reaches an intermediate frequency throughout central Europe, for example, at approx 0.550 in France and Germany, but frequency drops significantly in Southern Europe reaching a low of 0.134 in Greece. Similar frequencies exist in Eastern Europe and central Asia, for example 0.118 in the Russian Caucasus, 0.100 in the Burusho of Pakistan (Bersaglieri et al 2004) and the Middle East, 0.044 in Iran.

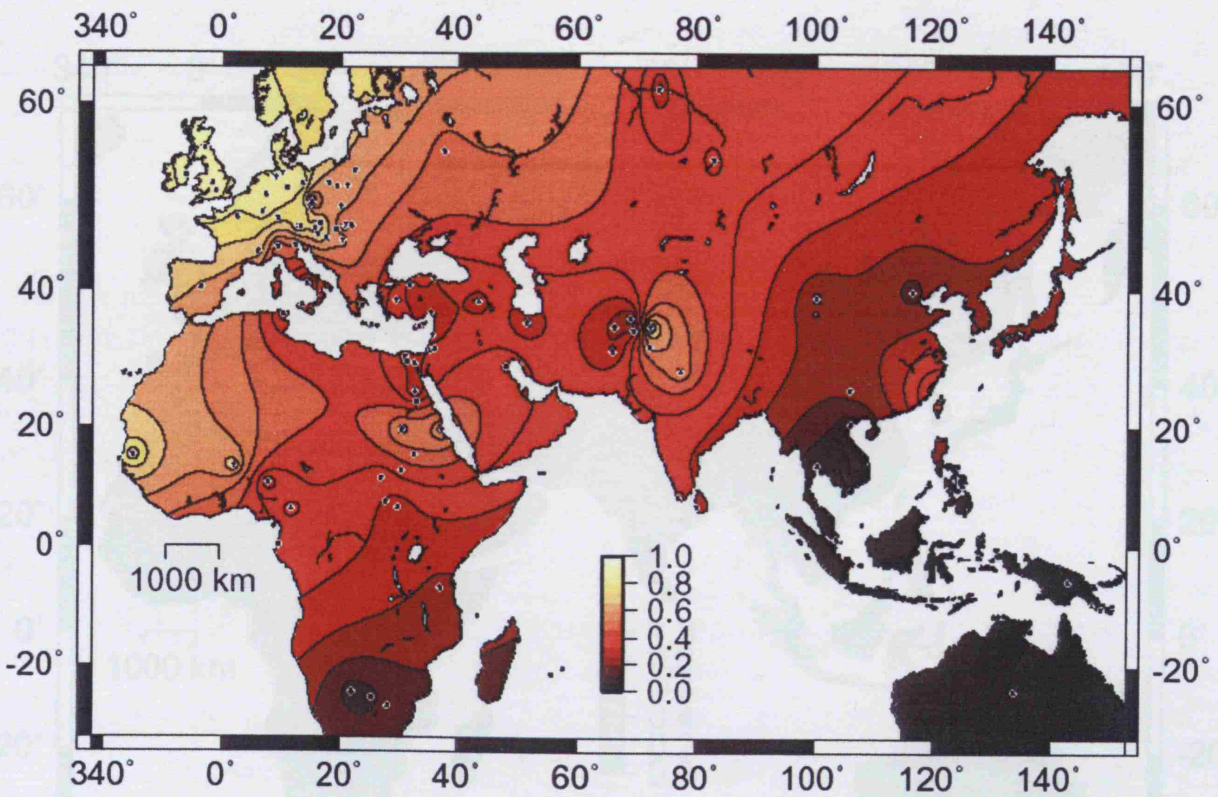


Fig. 5.2 An interpolated frequency map to show Old World distributions of lactase persistence

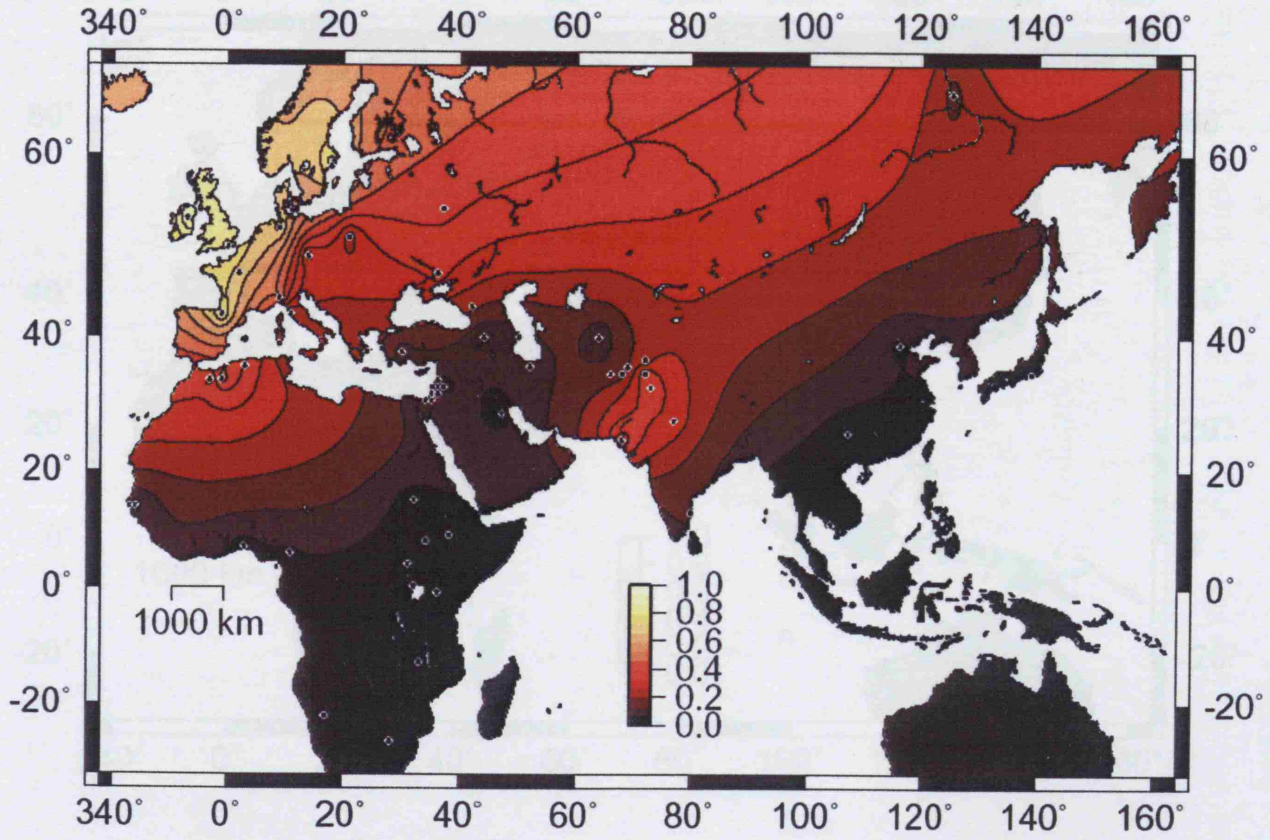


*Fig. 5.3 An inter-polated map showing the frequencies of lactase persistence as predicted by the –13.9kb*T allele*

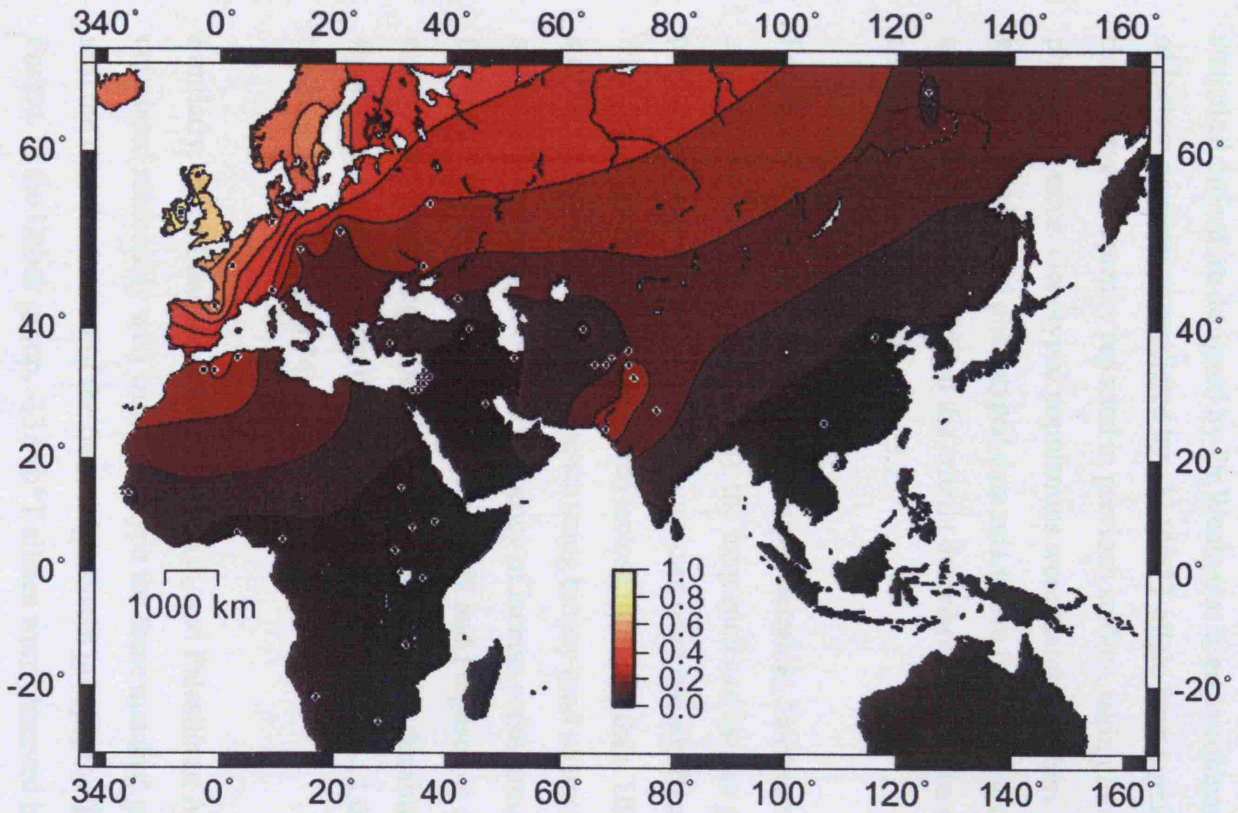


Figure 5.4 An inter-polated frequency map
*to show the Old World distribution of frequencies of –13.9kb*T allele*

*5.3.4 Can the –13.9kb*T allele predict lactase persistence frequency in Eurasia?*

The maps suggest that the T allele correlates well with lactase persistence in Western Europe, but that this association may break up further East. Using the TCGA population groups where full genotype information was known, the statistical procedure designed by Dr.Weale was used to compare the frequency of lactase persistence predicted by –13.9kb*T allele for a population group with the observed frequency reported in previous studies, taking into account a phenotype error. Genotyped populations were matched, where possible, with previously published phenotyped data sets where the ethnic group was the same, and where the study fulfilled the criteria described in 5.2. The results of the comparisons are shown on table 5.2.

For all of the European population groups (Germans, Northern French, Greeks and Irish as described in chapter 4), the frequencies of lactase persistence predicted using –13.9kb*T allele frequencies were not significantly different from that expected from the reported lactose tolerance data. UK English subjects tested for lactase persistence using biopsy (and so unsuitable for the statistical procedure) showed a frequency of lactase persistence ranging between 0.84 and 0.95 (Pena et al 1973, Ho et al 1982 and Ferguson et al 1984). Taking into account those lactase persistence is controlled by a dominant allele, these figures are consistent with the observed frequency of the –13.9kb*T allele in this population group (0.74).

Similarly, in the Iranians, Ashkenazi Jewish, and Palestinian Arabs, genotype correlated reasonably well with phenotype for these matched groups. This was also the case for two out of the three Afghanistan groups, the Tadjik and the Pashtu. In the Uzbek group, -13.9kb*T alleles were observed in the TCGA sample set at a frequency of 0.079 whereas no evidence of lactase persistence was found in the phenotyped matched group (Rahimi et al 1976). Despite a good correlation for the Palestinian Arabs, the two other Middle Eastern populations showed significant differences ($p = < 0.001$) between phenotype

predicted by –13.9kb*T allele and reported phenotype (Hijazi et al 1982 and Sanae et al 2003).

Most interestingly, the Turkish population group from the Anatolian plains (compared with a group from the central Anatolian region in Turkey for which phenotypic data was available) also showed significant differences between the –13.9kb*T predicted phenotype and reported phenotype. However, when the –22kb*A allele was used instead, the difference observed was not statistically significant ($p = 0.348$).

5.3.5 Association between the alleles associated with lactase persistence

Using PHASE software to establish the most likely haplotypes for each population group, as before, the –22kb*A and –13.9kb*T alleles associated completely with 5579*C. In some populations, the –22kb*A allele was found without the T allele. The ACC haplotype, (-22kb*A, -13.9kb*C, 5579*C) was seen in the Tadjik ($n = 1$), the Uzbek ($n = 3$), the Turks ($n = 5$), the Assyrians ($n=1$) and the Greeks ($n = 2$). The highest frequency of ACC haplotypes, 0.0761, was observed in the Ukrainian population ($n = 7$).

A comparison of the allele counts for 5579kb*C alleles between the Israeli Arabs and Israeli Bedouin showed a slight difference between the groups but was not statistically significant (**$p = 0.196$ – *Fisher's Exact Test***)

	*C-13.9kb*T Polymorphism data set*			*Published data set of phenotyped individuals*					
Population Group	**Number of chromosomes**	**Number of – 13.9kb*T alleles**	**Number of – 13.9kb*C alleles**	**Number of Individuals**	**Number of observed lactose digesters**	**Number of observed lactose maldigesters**	**Test method**	**Reference**	***P-value (to five decimal places)***
Germans	30	33	27	56	51	5	H	Flatz et al 1986	0.05554
Northern French	48	20	28	62	48	14	H	Cuddenac et al 1982	0.21660
Greeks	82	11	71	600	332	268	G	Kanaghinis et al 1974	**0.0002**
Ashkenazi Jews[a]	96	8	88	32	10	22	G	Leichter et al 1971	0.48626
Anatolian Turks	99	3	96	104	30	74	H	Flatz et al 1987	**0.01884**
Afghanistan – Pashtun	16	2	14	71	15	56	G	Rahimi et al 1976	0.35454
Afghanistan – Tadjik	98	10	88	79	14	65	G	Rahimi et al 1976	0.24954
Afghanistan – Uzbek	76	6	70	16	0	16	G	Rahimi et al 1976	**0.04482**
Iranians	90	4	86	40	7	33	G	Sadre et al 1979	1.12814
Kuwait	31	0	28	70	37	33	H	Sanae et al 2003	**0.00032**
Jordanian Bedouin	40	3	37	162	123	39	H	Hijazi et al 1982	**0.00000**
Palestinian Arabs	34	1	33	148	37	111	H	Hijazi et al 1982	0.26958

Table 5.2 Comparisons with published lactose digester frequencies in matching populations, taking into account sampling error.
Key: G = Blood Glucose Test, H = Breath Hydrogen Test, P-value = result of test described in 2.7 and 5.2, except where indicated. Significant p values are shown in bold. 'a' indicates the Ashkenazi Jews compared in the phenotyped population were from Canada, whereas those in the genotyped population were from Eastern Europe.

5.4 Discussion

If, as is hypothesised in this thesis, the lactase persistence phenotype arose in Neolithic pastoralists or early farmers, the historic movement of peoples can partly explain the modern breadth of observed lactase persistent chromosomes. If the trait emerged amongst the farmers of the fertile crescent, lactase persistent chromosomes could have been brought from the Near and Middle East to Europe and India in accordance with the traditionally held 'Wave of Advance' model (Ammerman and Cavalli-Sforza 1973). However, conversely, the clinal pattern of lactase persistence frequency operates in the reverse direction, reaching its highest frequency in Northern-Europe and declining towards Southeast Europe. This observation is broadly true for the –13.9kb*T allele as well.

The observed distribution of alleles associating with lactase persistence in the Eurasian data set suggests several possible interpretations of the global distribution of the lactase persistence trait and –13.9kb*T frequency. The data presented in this chapter suggests –13.9kb*T alleles are absent, or occur only at very low frequency in many areas known to have high numbers of lactase persistent individuals, as was the case in sub-Saharan Africa. Although–13.9kb*T allele frequencies were able to predict lactase persistence frequency in Western Europe, Iran, amongst the Ashkenazi Jews, in regions of Afghanistan and in the Palestinian Arabs, the allelic frequency was too low explain previously observed frequencies of the trait in the Kuwaitis, Anatolian Turks and Jordanian Bedouin. This suggests that lactase persistent chromosomes that do not carry the –13.9kb*T allele may exist elsewhere in Eurasia.

It is possible that, given the clinal direction of lactase persistence and also –13.9kb*T, the trait both emerged and reached high frequency amongst Northern-European groups, possibly even during the Mesolithic, and back-migration took lactase persistent chromosomes outwards into Europe. Under this scenario, the

disassociation between the derived *MCM6* alleles (Enattah et al 2002) and lactase persistence in central Eurasia would need to be explained by subsequent recombination and drift, or a separate mutational event leading to the independent evolution of lactase persistence outside of Europe, as is likely to have occurred in sub-Saharan Africa

Alternatively, the –13.9kb*T allele could have emerged somewhere in the Neolithic migratory route from the fertile-crescent to North-West Europe, perhaps in central Europe or even parts of Asia. Even if the highest frequencies of lactase persistence occur in North-West Europe, the earliest lactase persistent individuals may not have been from this region. Lactase persistent chromosomes carrying –13.9kb*T may have reached the very high frequencies observed today in North-West Europe due to the effects of strong drift and/or selection. Under a Wave of Advance model, recent computer simulations have shown this to be possible (Edmonds et al., 2004). This scenario will be investigated in more detail in chapter 6.

An order to gain further insight into the possible origins of lactase persistence and –13.9kb*T allele, one approach was to track the –13.9kb*T allele and a putative progenitor A haplotype. The G-A SNP at –22kb, and the 5579*C allele which defines the A Haplotype were used as extra markers for this purpose. The data in Chapter 3 indicates that both alleles are associated with lactase persistence but also that they both predate the –13.9kbT allele. The theoretical 'progenitor' haplotype, that is, a haplotype that might carry a causative mutation prior to the emergence of –13.9kb*T allele, was defined as the condensed SNP Haplotype ACC. If this haplotype had been observed at high frequency, possibly in a population with some –13.9kb*T carrying chromosomes at low frequency, then from a phylogeographic perspective it might be hypothesised that the –13.9kb*T allele arose in this location and diffused out. Further lactose tolerance tests targeted in such a place might reveal lactose tolerant individuals with –22kb*A, but without the –13.9kb*T allele.

Unfortunately, the –22kb*A and –13.9kb*T SNPs were extremely tightly associated in most populations, and so the ACC haplotype was not as informative for investigating the geographic origin of the –13.9kb*T allele as was hoped. Even so, several interesting observations were possible. The ACC haplotype was observed widely, albeit at low frequency, in the Tadjik, Uzbek, and Ukrainian, Assyrian, Greek and Turkish groups. Bersaglieri et al also noted higher frequencies of –22kb*A allele in Pakistan (where, conversely, –13.9kb*T alleles were found without the –22kb*A, in the Sindhi), the Uygur in China and the Israeli Bedouin groups (Bersaglieri et al 2004). The distribution of the –22kb*A allele without the T allele, in particular, strongly the hypothesis that the T allele (and lactase persistence) emerged in the Fertile Crescent rather than in Mesolithic or Paleolithic Europe. However, if the –13.9kb*T allele was subject to strong selection mainly in northern Europe then this might obscure phylogeographic signals of a northern European origin.

Of particular interest is the fact that the Anatolian Turkish sample had a frequency of 0.051 for the ACC haplotype, and, in this group it was not possible to predict lactase persistence frequencies using the –13.9kb*T allele though it was possible using the –22kb*A. Given the great importance of this region, both as a part of the Fertile Crescent where the earliest evidence of agriculture was found, and also as the main route for dispersing agricultural practice into Eurasia. It seems that this area, and similar population groups, should be researched further. It may be the case that lactose tolerance testing may show lactase persistent chromosomes with the condensed SNP haplotype of ACC and ACT exist in this group. If that were the case, it would strongly support a single causal mutation, with some allelic association dominating the genetic profile of lactase persistent individuals in Western Eurasia, and perhaps different associations occurring in the populations of the Fertile Crescent. However, without more extensive typing of polymorphisms and known lactase persistence status in this region, conclusive evidence cannot be established for a single mutational event or genetic heterogeneity of causes for lactase persistence.

Future project work looking at other alleles that sub-divide the A Haplotype, such as the InDel in intron 1, might enable a more resolved survey of Central Europe where the association between the derived *MCM6* alleles and lactase persistence breaks down. Outside of Europe, and particularly in the region covering the Fertile Crescent, it seems probable that other, as yet unidentified alleles associate with lactase persistence. In this context, it is of interest to note that the 5579*C allele defining the A Haplotype in Europe is not markedly different between the Bedouin in Israel, known to have a long-standing history of fresh milking drinking and a high frequency of lactase persistence (Cook and Al-Torki 1975), and the Israeli Arab group, where lactase persistence is rarer. The 5579*C allele is not particularly frequent in either group, being lower in the Bedouin. It may, therefore, be the case that in the Middle East, as in sub-Saharan Africa (see Chapter 4), lactase persistent chromosomes exist that do not carry the –22kb*A or the –13.9kb*T alleles, and that are also not on the background of an A Haplotype.

Chapter Six

Evidence for selection of lactase persistence

6.1 Introduction

The search to find evidence of natural selection in modern human populations has generated several interesting studies (for example, as discussed in the introdiction, Hughes and Nei 1988, Rooney and Zhang 1999, Slatkin and Bertorelle 2001, Zhang et al 2002). As more potential examples are reported, it is clear that although these studies provide evidence of selection, they also raise questions about its nature and processes.

In 2001, Wiuf used three microsatellite loci and, investigating the frequency of ΔF508 alleles, generated evidence to suggest that natural selection had occurred at the cystic fibrosis locus in Europe. However, it was not possible to determine whether the selection positively favoured ΔF508 heterozygotes or was indicative of negative selection against other alleles. By analogy, it is possible that selective pressures in lactase are negative, acting against non-persistent individuals in certain populations rather than favouring the persistence trait.

The high frequency of sickle cell anaemia in some populations has often been quoted as an example of natural selection in modern humans (Haldane 1949). A study by Pagnier et al (1984) investigated haplotypes defined by 11 loci distributed inside a 60kb region within the B-globin gene cluster. Looking specifically at chromosomes carrying the sickle-cell trait, they found that in Benin and Algeria, the haplotype associated with the sickle cell trait was identical. However, in the Central African Republic and in Senegal respectively, the trait associated with different haplotypes. They concluded that the sickle cell allele, existing at high frequencies in each geographically distinct region, emerged independently on separate haplotypic backgrounds and each of these was subject to selection. A more recent study confirmed evidence of independent evolution and negative selection due to malaria (Tishkoff et al 2001). Ths study also estimated the date of the

emergence of the sickle cell alleles as concurrent with the emergence of agriculture, coinciding with the development of irrigation techniques that led to stagnant water supplies, which favour the malarial parasite. This creates an interesting precedent, which, given the findings of chapters 4 and 5, may be relevant for lactase.

The Duffy Blood gene group locus (FY) has three common alleles, FY*A, FY*B and FY*O, with frequencies that are strongly differentiated by geographical region. The FY*O allele, which is thought to confer some protection against malaria, has reached near-fixation in the Hausa of Cameroon, and shows evidence of directional selection (Hamblin et al 2002). The authors also found evidence of high frequency at the FY*A allele in non-African populations, but it was not possible to explain this by a simple model of selection. Hamblin and colleagues suggested that the complex patterns of variation they observed could best be explained by the interaction of selection with a complex demographic history and/or selective pressures acting on more than one allele (Hamblin et al 2004).

The examples described above suggest that, although there are many candidate gene regions and variants, which have reached current frequencies due to natural selection, the nature, expression and interpretation of selection signatures is complex. The unusual distribution of lactase persistence phenotype has, for a long time, been considered as an example of the results of natural selection, (Simoons 1970, 1978, McCrackern 1971, Aoki 1986, Holden and Mace 1997), with the added interest of an interplay between human culture and human biology, the former creating the evolutionary conditions for the latter. Although intuitively the correlation of a high frequency of the lactase persistence phenotpye and a history of milking suggests a selection pressure in favour of those able to drink fresh milk (for review, see Swallow 2003) it is important to show evidence for this statistically, and to estimate the likely scale and timing of such a demographic process.

Recently, Bersaglieri et al (2004) published a study designed to determine whether selection had occurred favouring lactase persistence. They compared F_{ST} levels for

the C–13.9kb*T polymorphism against (presumed) neutral SNPs in other regions of the genome in the same samples. A series of Coriell samples from 3 population groups, European Americans, African Americans and East Asians were investigated for allelic frequency of 28,400 markers to generate a null distribution of F_{ST}. Looking at allelic frequencies of the –13.9kb*T and –22kb*A alleles in the same individuals, Bersaglieri and colleagues found F_{st} levels of 0.53 in the comparisons between Europe and East Asia, and Europe and Africa, in each case (2004). This was higher than that of 99.9% of the presumed neutral SNPs typed in the same individuals, emphasising the unusual allelic distribution of the derived *MCM6* alleles.

The authors also used a modified method of Sabeti et al (2001) (see also section 1.1.2.3), and a statistic called 'p_{excess}' to test for inter-allelic diversity between haplotypes carrying the –13.9kb*T allele and those carrying the –13.9kb*C allele. They showed evidence for significantly less diversity and longer haplotypes (>1MB) in –13.9kb*T carrying chromosomes, which existed at a high frequency (77% approximately). This observation was considered consistent with the notion that –13.9kb*T carrying chromosomes rose to high frequency in a population comparatively quickly, with insufficient time for recombination to disrupt these long haplotypes. This interpretation of the data strongly suggests a historic selection pressure.

However, as Bersaglieri and colleagues discuss in their conclusion, it is possible that dominant suppression of recombination due to an allele on the –13.9kb*T carrying chrommosomes (which is not, in itself, connected to lactase persistence) may have affected the results (Bersaglieri et al 2004). Similarly, molecular causes relating to LD in that region might confound selection tests where extended haplotype blocks are the measure of inter-allelic diversity.

Using a dating technique based on the decay of haplotypes at both the 3' and 5' end (Reich and Goldstein 1998, Stephens et al 1998), Bersaglieri and colleagues

estimated a time of between 2,188 and 20,650 years ago, consistent with the development of agriculture, for the emergence of the–13.9kb*T carrying chromosomes. Using this estimated date, the authors proposed a selection co-efficient of between 0.014 and 0.15 based on data from the CEPH sample they used, and 0.09 and 0.19 based on data from a Scandinavian sample. Bersaglieri and colleagues commented that this level of selection operating on lactase is comparable with that observed at the *G6PD* locus, affecting phenotype of the sickle-cell trait protective against malaria, with a reported a selection coefficient of 0.02-0.05 (Tishkoff et al 2001).

While this work was in progress, the Syssiphos program used in this thesis was being developed (see also 2.6.6). Syssiphos uses data from closely linked microsatellite loci to identify possible signatures of selection, and so provides an alternative approach to that used by Bersaglieri and colleagues for investigating selection for the lactase persistence trait.

Whether or not –13.9kb*T is causative of lactase persistence, there is clearly a strong correlation between lactase persistence and –13.9kb*T allele in Europe; selection tests can therefore justifiably use this polymorphism to identify lactase persistent chromosomes in a sample set from that region, even if the –13.9kb C to T transition has no functional role in lactase persistence itself. However, outside of Europe, where –13.9kb*T allele is infrequently found, this is not the case. Since –13.9kb*T occurred on the background of a common haplotype, A, which is characterised by 5579*C allele and is more frequent outside Europe than –13.9kb*T, this polymorphism was also examined.

The key aim of this chapter is to investigate whether a historic selection pressure has brought the –13.9kb*T allele to high frequency in Europe, and, using microsatellite-based dating techniques, to estimate the age of this allele. This chapter reviews a new method for identifying selection, (see also 2.6.6) and tests the effects of altering a series of variables to evaluate its sensitivity to model

parameters. Finally, in those populations where the –13.9kb*T allele is not present but which have a history of pastoralism, a possible selection signature for the 5579*C allele is investigated.

6.2 Methods

6.2.1 Samples and practical methodology

The methods used in this chapter are described in section 2.4-2.6. Five microsatellite loci and three SNP loci were investigated. Figure 6.1 and table 6.1 shows the chromosomal positions of the loci under investiation, positions relative to the C–13.9kbT locus and recombination rates that were used as parametres for the Syssiphos program. Phase was established by pedigree analysis in nine population groups comprised of families as described previously in 2.5.2 and chapter 3.

6.2.2 Statistical analysis – descriptive statistics

Summary statistics were estimated as described in section 2.6, with mean, modes, medians and variance of microsatellite repeat numbers observed for each microsatellite locus calculated in excel. The following descriptive statistics were performed in Arlequin (Schneider et al 2000): Exact Test of Population Differentiation (Raymond and Russet 1995), AMOVA (Wier and Cockerham 1984, Excoffier et al 1992, Wier 1996), and pairwise comparisons of populations based on microsatellite data using a sum of squared distance test, R_{ST}, (Slatkin 1995).

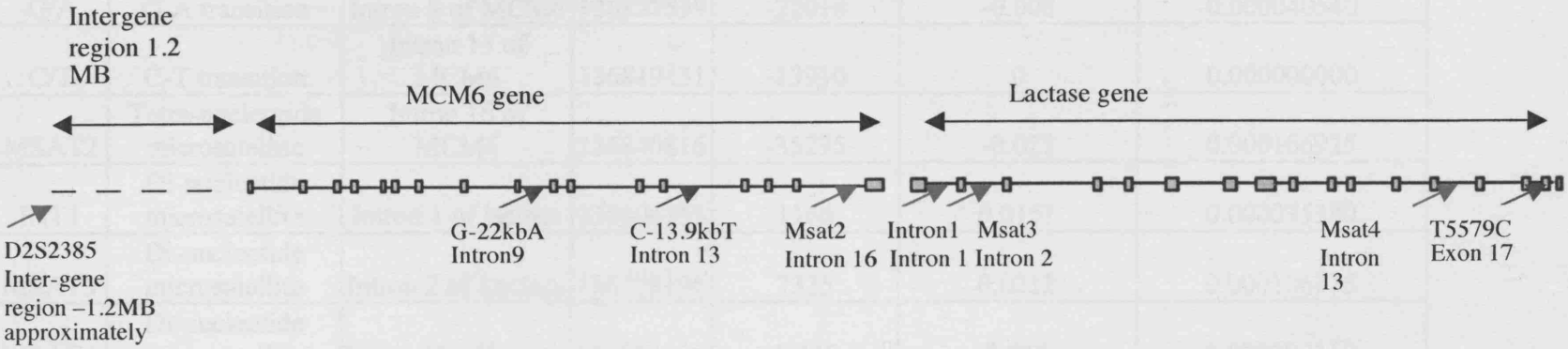


Fig 6.1. A digram to show the positions of the SNPs and microsatellites referred to in this chapter.
The blue arrows denote SNP loci, the red arrows microsatellite loci. The labels given to each refer to the actual names of the loci as used in this thesis, and below is the approximate position of each in the gene.

Marker name	Class of Polymorphism	Location	Position no[35]	Distance from the start of lactase in bp	Position relative to C-13.9kbT (Mb)	Recombination rate per generation[36]
D2S2385	Di-nucleotide microsatellite	Intergene region approx. 1.2MB upstream of lactase	137972683	-1E+06	-1.153	0.005866260
G/A	G-A transition	Intron 9 of MCM6	136827539	-22018	-0.008	0.000040540
C/T	C-T transition	Intron 13 of MCM6	136819431	-13910	0	0.000000000
MSAT2	Tetra-nucleotide microsatellite	Intron 16 of MCM6	136840816	-35295	-0.021	0.000106925
INT1	Di-nucleotide microsatellite	Intron 1 of lactase	136804355	1166	0.0151	0.000075380
MSAT3	Di-nucleotide microsatellite	Intron 2 of Lactase	136798196	7325	0.0212	0.000106175
MSAT4	Di-nucleotide microsatellite	Intron 13 of lactase	136763409	42112	0.056	0.000280110
Exon 17	C-T transition	Exon 17 of lactase gene	136756830	48691	0.0626	0.000313005

Table 6.1 Microsatellite descriptions, locations and recombination fractions

[35] Position given in base pairs from the Human Genome Browser (http://genome.cse.ucsc.edu/cgi-bin/hgGateway) July 2003 freeze

[36] Recombination rates calculated from sex-averaged Genethon and DeCode data, that is, assuming 0.5cM per Mb, as downloaded from the Human Genome Broswer web-site July 2003 freeze

6.2.3 *Syssiphos program*

The Syssiphos program[37] calculates the likelihood of obtaining a given data set (in this case a series of chromosomes made up of compound microsatellite haplotypes) given different values of population growth and selection. At the time of writing, the program was unpublished, and kindly made available by Dr.Michel Stumpf. This being the case, much of the work using Syssiphos was experimental, and part of this chapter reports the effects of trialling different parameters values in the program.

An unbound, length-dependent stepwise mutation model is assumed (Stumpf and Goldstein 2001), and in addition, recombination is modelled assuming an ancestral haplotype distribution that is the same as the present one. Coalescent trees are simulated given a user-specified maximum tree depth and microsatellite mutation rate (see section 2.6.6 for detailed description). The program enables the maximum likelihood microsatellite compound haplotypes on different SNP backgrounds to be compared. For analysis in the Syssiphos program, populations were grouped into a series of different combinations, in order to increase the data for each set of analyses. These amalgamated groups would also reflect the geographical and historical associations of the populations (table 6.2).

Group	Populations included	Number of Chromosomes
Ashkenzi Jews	Ashkenzi Jews only	96
Ethiopia	Amharic Ethiopians only	119
Eurasia	Ireland, UK English, France, Germany, Ashkenazi Jewish and Armenian	411
Middle East	Armenia and Kuwait	116
Africa	Algeria and Ethiopia	140
Western Europe	Irish, UK English, French, German	227
The Whole World	Armenian Kuwaiti, Algerian, Ashkenazi Jewish, Irish, UK English, French, German, Amharic Ethiopian	579

Table 6.2 Family populations grouped by geographic region

[37] The program was written by and kindly made available by Dr. Michel Stumpf

Three tranches of experiments using Syssiphos were undertaken. The first set of experiments investigated the effects of changing Syssiphos parameters. The second involved a 'rough check' to search for maximum likelihoods of the data under a wide range of different combinations of selection and growth, and the final tranche involved a fine-scale, in depth exploration of selection and growth parameter space where maximum likelihoods were found in the 'rough check'.

For every Syssiphos simulation, an output file was generated, which comprised a list of the likelihood of each of the values of growth and selection within the set parameters. A script written in the R statistical programming enviroment[38] (Appendix A2) was used to visualise the Syssiphos output as either a two-dimensional or a three-dimensional representations of the likelihood surface. The code incorporates a function that removes the 'tails' of the maximum likelihood output, such that less probable likelihoods are not shown on the graphs.

[38] The 'R' Code, first described in Caldwell (2005), was kindly provided by Dr.Mark Thomas

An example of a typical two-dimensional graph is shown below where the coloured diagonal in the graph represents the values of likelihood. An example of a three dimensional graph is shown in figure 6.9 – 6.12, where the likelihood values are shown as a third axis, or 'peak'.

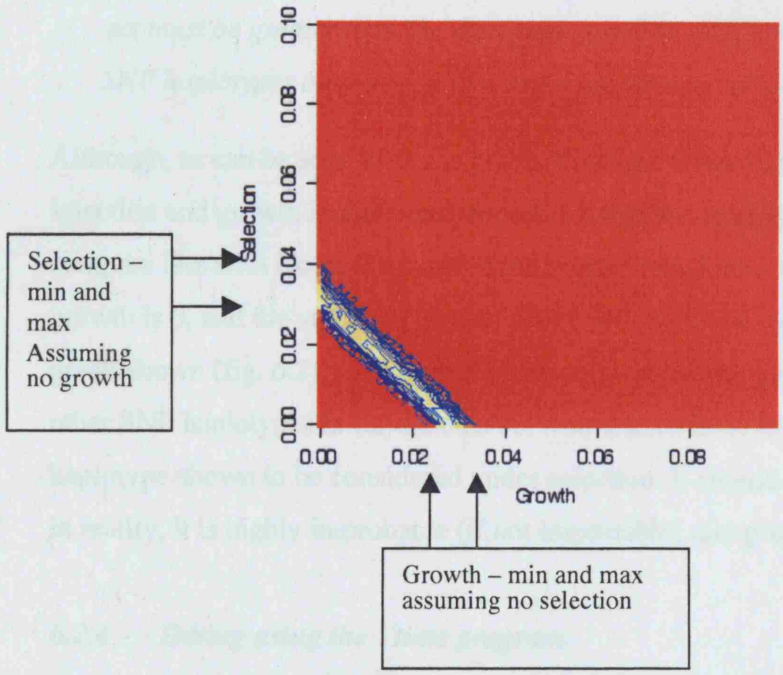


Fig 6.2 An example of a low- resolution two-dimensional image using output generated by the Syssiphos program. *The minimum and maximum possible values of selection and growth were read off each graph as shown by the arrows in bold. In this example, the readings are of 0.0275 and 0.04 for selection (assuming growth is 0) and 0.0023 and 0.033 for growth (assuming selection is 0) respectively. The blue contour lines indicate the values of likelihood generated by Syssiphos for the data, assuming the levels of growth and selection at that point on the graph. The red portion of the graph indicates areas where the likelihood value was more than 2 units below the maximum likelihood value obtained.*

Once the minimum and maximum values for selection and growth had been recorded for each SNP haplotype in each population set, the results were reviewed and a selection signature was interpreted by the following rule:

> *The minimum possible selection value for a SNP haplotype in one population set must be greater than the maximum possible selection value(s) for all other SNP haplotypes observed in that same population set assuming the same value*

Although, as can be seen from the graph, there are numerous combinations of selection and growth in different proportions that are equally likely, in order to compare like with like it is easiest in the graph to read the values of selection where growth is 0, and the values of growth where selection is 0. So, from the example graph shown (fig. 6.2), the maximum selection (assuming growth is 0) for all the other SNP haplotypes in for the data set would need to be less than 0.0275 for the haplotype shown to be considered under selection. It should, however, be noted that in reality, it is highly improbable (if not impossible) that growth should ever be 0.

6.2.4 Dating using the Ytime program

As described in 2.7, Ytime[39] was used to calculate the most likely T_{MRCA} (Time to Most Recent Common Ancestor) for a set of chromosomes based on microsatellite data. A mutation rate of 0.0012 was assumed as before (Weber and Wong 1993) and a simple stepwise mutation model assuming a star-genealogy was used. Confidence intervals of 0.025 and 0.975 were established for the date. T_{MRCA} in number of generations was also estimated in Excel, using the basic equation $T_{MRCA} = ASD/\mu$, where ASD = average square difference between the ancestral chromosome (here assumed to be the most common haplotype) and μ = mutation rate. Generation time was assumed to be 25 years.

[39] Ytime, written by Dr.Mike Weale, is available from the TCGA web-site, www.tcga.com/statistics

6.2 Results

6.3.1 Raw Data description

584 chromosomes were resolved for the compound SNP and microsatellite haplotypes, by following the pattern of inheritance in the families. Table 6.3 below summarises, for some of the amalgamated groups of populations, the number of different compound haplotypes observed. The greatest was, as might be expected, in the Ethiopians, (53). Table 6.4 summarises the mean, mode, range and variance of each of the microsatellite loci. The greatest range of repeats for any microsatellite was found in intron1 (4 – 23) the smallest range of repeats was for MSAT2, the tetranucleotide (4 - 8).

Population group	SNP haplotype			Total n of chromosomes
	ATC	GCC	GCT	
Ethiopians	0 (n = 0)	14 (n = 25)	39 (n = 94)	119
Middle Eastern	1 (n = 1)	15 (n = 30)	31 (n = 85)	116
Western European	13 (n = 159)	11 (n = 22)	14 (n = 46)	227

Table 6.3. A table to show the number of microsatellite haplotypes associated with each of the core SNP haplotypes, for three of the groups. *The total number of SNP haplotypes found in each group is shown in brackets*

	D2S2385				MSAT2				INTRON 1				MSAT3				MSAT4			
Population	Min	Max	Mode	Mean	Min	Max	Mode	Mean	Min	Max	Mode	Mean	Min	Max	Mode	Mean	Min	Max	Mode	Mean
Algeria (n = 21)	14	25	26	22.1	5	7	6	6.2	5	19	17	8.8	14	17	16	15.8	11	12	11	11.2
Armenia (n = 88)	11	27	20	19.7	5	8	6	6.5	7	20	10	10.6	9	18	17	16.1	11	12	11	11.1
Ashkenazi Jewish (n = 96)	14	26	22	21.1	6	7	7	6.5	7	19	10	9.7	14	18	17	16.3	11	20	11	11.3
Amharic Ethiopian (n = 119)	14	26	23	20.8	4	7	7	6.5	5	23	10	11.0	12	19	17	16.5	9	18	11	11.1
UK – British (n = 64)	14	26	20,22	21.8	6	7	6	6.1	6	16	7	7.7	15	18	15	15.5	10	12	11	11.1
German (n = 60)	14	26	22	21.6	6	7	6	6.3	7	18	7	8.6	14	18	16	15.6	11	13	11	11.1
Irish (n = 65)	14	27	22	22.0	6	7	6	6.0	4	22	7	7.2	14	17	15	14.9	11	12	11	11.0
French (n = 38)	14	27	22	21.4	6	7	6	6.3	7	11	7	8.2	14	17	15	15.7	11	12	11	11.0
Kuwaiti (n = 28)	19	25	22	21.7	5	7	6	6.3	5	19	10	9.5	15	17	16,17	16.3	10	16	11	11.2
Total:	11	27	22	21.14	4	8	6	6.34	4	23	7	9.37	9	19	16	15.93	9	18	11	11.11
Variance for total data set:	10.22				0.253				11.09				1.283				0.465			

Table 6.4 Mean, Mode and ranges of repeat number observed for each microsatellite locus in each population group, and total variance in microsatellite repeat for the data set as a whole shown for each microsatellite locus

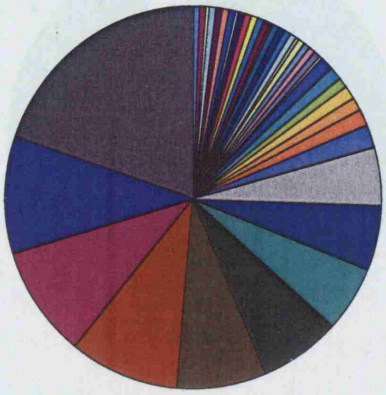
6.3.1.1 Microsatellite variation within the three condensed SNP haplotype

Figures 6.3-6.8 show the microsatellite haplotype frequencies for the whole data set divided up by condensed SNP haplotype. The most distant marker, D2S2385, was first included (figs 6.3-6.5), but inspection of the family data indicated that it showed very little allelic association with the other loci. When the D2S2385 locus was removed, 145/174 **ATC** chromosomes could be seen to carry a common microsatellite haplotype of **6, 7, 15, 11.** When the ATC-carrying chromosomes for each Western European population are considered in isolation, there appears to be a slightly higher number of microsatellite haplotypes associated with the ATC SNP haplotype proportionate to the total number of ATC chromosomes in Ireland (n=9/62) and England (n=7/47), as distinct from France (n=1/17) and Germany (n=2/33).

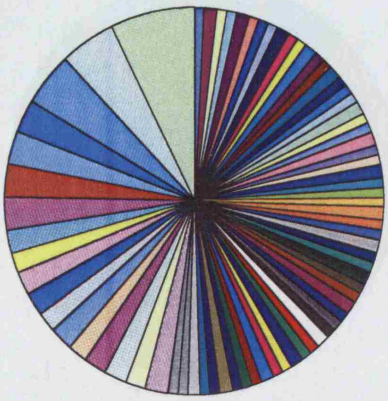
Table 6.6 summarises the variance at each microsatellite loci for each of the three core SNP haplotypes for the total data set. It appears that there is relatively less difference in variance between the SNP haplotypes for the D2S2385 locus compared with the others.

	SNP haplotype		
Microsatellie Locus	**ATC**	**GCC**	**GCT**
MSAT2	0.0057	0.1174	0.2831
MSAT3	0.0974	1.1707	1.0176
MSAT4	0.0335	0.2317	0.8039
INT1	1.5373	15.0836	10.3560
D2S2385	7.8129	11.3202	10.7274

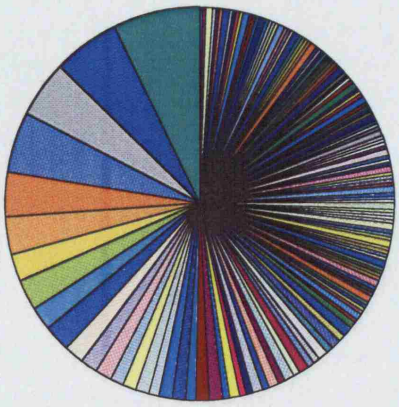
Table 6.5 - A table to show the variance for each locus when the microsatellite data is grouped by core SNP haplotype



ATC

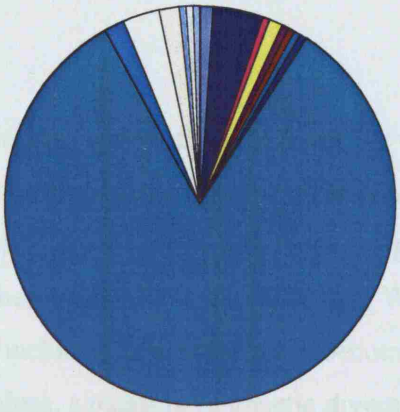


GCC

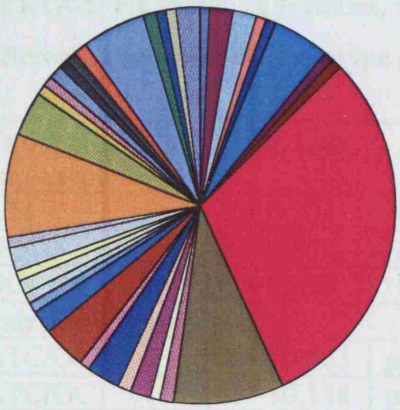


GCT

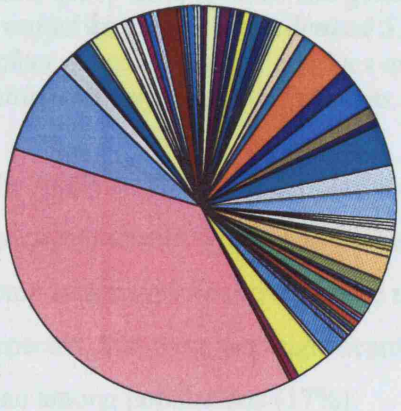
Figs 6.3, 6.4 and 6.5. Three pie charts to show microsatellite haplotype frequencies for each of the SNP haplotypes, taking into account all five microsatellite loci. *Each colured segment of the pie chart represents a different microsatellite haplotype, with the size of the segments representing the proportion of the total number of haplotypes observed.*



ATC



GCC



GCT

Figs 6.6, 6.7 and 6.8. Three pie charts to show microsatellite haplotype frequencies for each of the SNP haplotypes, taking into account four of the microsatellite loci. *Each coloured segment of the pie chart represents a different microsatellite haplotype, with the size of the segments representing the proportion of the total number of haplotypes observed*

The pie charts reflect the diversity within each condensed SNP haplotype, as determine by the microsatellite data. What can be seen clearly is that, when just four loci are considered, there is significantly less diversity in the ATC haplotype when compared to the other two. When the distant microsatellite marker D2S2385 is included, this observation becomes less clear. The table below shows the *h* values, a measure of genetic diversity, for each condensed SNP haplotype as determined by the microsatellite data. Comparisons between the condensed haplotypes are shown as z-values, which can be used to show significant differences between the haplotype groups.

Condensed SNP haplotype	**Four Loci**		**Five Loci**	
ATC	h = 0.304	+/- 0.046	h = 0.900	+/- 0.012
GCC	h = 0.889	+/- 0.024	h = 0.988	+/- 0.004
GCT	h = 0.850	+/- 0.020	h = 0.983	+/- 0.003
Pairwise comparisons				
ATC/GCC	z = 10.863	**p = < 0.01**	z = 0.733	p = 0.45
ATC/GCT	z = 10.138	**p = < 0.01**	z = 0.692	p = 0.48
GCC/GCT	z = 1.379	p = 0.16	z = 1.118	p = 0.25

Table 6.6 A table to show the genetic diversity of microsatellite haplotypes when grouped by the three condensed SNP haplotypes, ATC, GCC and GCT. *h values reflect genetic diversity, z values are used for pairwise comparisons and statistically significant differences are shown in bold.*

The Analysis of MOlecular Variance (AMOVA) test was used to investigate the genetic structures within and between populations, taking into account the allelic composition and frequency of the microsatellite data. The results showed, as was expected, that there was significantly more variation within populations (83%) than among populations (17%).

Source of variation	Degrees of Freedom	Sum of Squares	Variance components	Percentage of variation
Among Populations	8	1227.856	2.27730 Va	16.98
Within Populations	570	6347.691	11.13630 Vb	83.02

Table 6.7 A table to show the AMOVA results for the microsatellite data (Wier & Cockerham 1984, Excoffier et al 1992, Wier 1996. Performed in Arlequin (Schieder et al 2000)

Table 6.8 shows the R_{ST} values for a series of pairwise comparisons between population groups, with the significant results shown in bold. Surprisingly, the R_{ST} values for many of the pairwise comparisons show significant differences between the population groups, even between the European groups. This is most notable in Ireland, which, the raw data suggests, has very little microsatellite diversity. There were fewer significant pairwise differences between the Kuwaiti group and other population groups, probably due to its comparatively small sample size.

Similarly, significant values were observed for many of the pairwise comparisons that were generated by the Exact Test of Population Differentiation shown in table 6.9.

	Algerian	Armenian	Ashkenazi	British	Ethiopian	French	German	Irish	Kuwait
Algerian	/	0.04505+-0.0203	0.09910+-0.0252	0.04505+-0.0203	0.00901+-0.0091	0.36036+-0.0429	0.72072+-0.0384	0.00000+-0.0000	0.43243+-0.0265
Armenian	0.08297	/	0.03604+-0.0201	0.00000+-0.0000	0.23423+-0.0473	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.18018+-0.0407
Ashkenazi	0.03217	0.02607	/	0.00000+-0.0000	0.02703+-0.0139	0.00000+-0.0000	0.00901+-0.0091	0.00000+-0.0000	0.82883+-0.0446
British	0.0627	**0.29036**	**0.24022**	/	0.00000+-0.0000	0.23423+-0.0562	0.02703+-0.0139	0.00901+-0.0091	0.00000+-0.0000
Ethiopian	**0.11066**	0.00111	**0.04743**	**0.29571**	/	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.06306+-0.0194
French	-0.00184	**0.20343**	**0.14728**	0.01166	**0.21527**	/	0.40541+-0.0339	0.00000+-0.0000	0.00901+-0.0091
German	-0.02729	**0.14487**	**0.0884**	**0.04181**	**0.17376**	-0.00565	/	0.00000+-0.0000	0.12613+-0.0242
Irish	**0.19559**	**0.38555**	**0.36768**	**0.05673**	**0.38177**	**0.16111**	**0.15574**	/	0.00000+-0.0000
Kuwait	-0.01832	0.01788	-0.01875	**0.16757**	0.03858	**0.07798**	0.03328	**0.29383**	/

Table 6.8 Comparisons of pairs of populations using Sum of Squared Distance analysis (R_{ST}) of microsatellite data. *Rst were calculated according to Slatkin (1995) assuming a stepwise model and implemented in Arlequin software (Schneider et al 2000). A probability for these values was also given. Comparisons giving a significant R_{ST} value, taking into account the error rate given by Arlequin, at <0.05 are shown in bold.*

	Algerian	Armenian	Ashkenazi	UK-English	Ethiopian	French	German	Irish	Kuwait
Algerian	/								
Armenian	0.07080+-0.0261	/							
Ashkenazi	0.09755+-0.0323	0.06000+-0.0133	/						
UK-English	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	/					
Ethiopian	**0.04000+-0.0153**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	/				
French	**0.00000+-0.0000**	0.20950+-0.0313	0.19910+-0.0405	**0.01330+-0.0099**	**0.00000+-0.0000**	/			
German	**0.00105+-0.0012**	**0.00145+-0.0016**	**0.02425+-0.0081**	0.27725+-0.0240	**0.00000+-0.0000**	0.75425+-0.0227	/		
Irish	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	/	
Kuwait	0.79540+-0.0141	**0.00860+-0.0044**	**0.00000+-0.0000**	**0.00000+-0.0000**	0.11100+-0.0754	**0.00000+-0.0000**	**0.00000+-0.0000**	**0.00000+-0.0000**	/

Table 6.9 A table to show the p-values and standard errors for the Exact Test of Population Differentiation using Microsatellite data. *Significant comparisons are shown in bold*

6.3.2 Testing the Syssiphos software

Before the compound SNP and microsatellite haplotypes could be analysed, the Syssiphos program itself was tested to determine whether altering the parameters affects the data, and, if so, how.

Overview of Syssiphos:

In order to estimate the likelihood of a the observed microsatellite diversity associated with a particular SNP allele at a given frequency, the program uses two raw data sets: one input file comprises all the microsatellite haplotypes observed in the population sample, and the other only includes those associated with a particular SNP haplotype or allele under investigation. The latter file provides the data upon which the main likelihood calculations are performed. The former file provides an estimate of the general microsatellite diversity in the population and is used to model recombination among different microsatellite and SNP loci. A third input file contains information on the frequency of the allele under investigation as well as other parameters used in the likelihood estimation.

In the present study, the SNP haplotypes described in chapter 3 and 6.2 were used, that is, ATC, GCC and GCT, where the first locus is the G-22kbA polymorphism, the second is the C-13.9kbT and the third is the T5579C locus which defines the A Haplotype.

From the data, the program generates a series of random trees and then calculates the likelihood of the data given these trees and other supplied parameter values under a range of combined values of growth and selection. This process is repeated for a wide variety of combinations of growth and selection, (for example, a selection coefficient of 0.001 and a growth coefficient of 0.001, a selection

coefficient of 0.002 and a growth coefficient of 0.001 etc) until a graph can be produced showing the likelihood surface for the range of values. Since Syssiphos is estimating the likelihood of a data set for different values of growth and selection (and other, non-variable parameters), a range of values for these parameters needs to be specified. Initially, these are guesses and can include values that achieve relatively low likelihood scores. However, later runs can be preformed using narrower ranges for selection and growth. Because signatures of selection are indistinguishable from signatures of population growth, the key aspects of the syssiphos outputs that were compared and reported in the following section are:

(1) Selection minimum - assuming that population growth is zero, the lowest value for selection that gives a likelihood of the data within 2 likelihood units of the maximum likelihood value.

(2) Selection maximum - assuming that population growth is zero, the highest value for selection that gives a likelihood of the data within 2 likelihood units of the maximum likelihood value.

(3) Growth minimum - assuming that selection is zero, the lowest value for population growth that gives a likelihood of the data within 2 likelihood units of the maximum likelihood value.

(4) Growth minimum - assuming that selection is zero, the highest value for population growth that gives a likelihood of the data within 2 likelihood units of the maximum likelihood value.

Syssiphos models selection and population growth under various parameter values of mutation rate, population size, maximum depth of coalescent tree. Some of these parameters, for example, recombination, are knowable, in that reliable

estimates are available. For effective population size, mutation rate and depth of tree, it was important to examine the effect on the simulations of changing these variables. The factors investigated were: the specified mutation rate, number of runs (of simulations), the maximum depth of coalescent tree and the effective population size. In each experiment, one variable was altered, and the outputs were compared.

6.3.2.1 Mutation rate

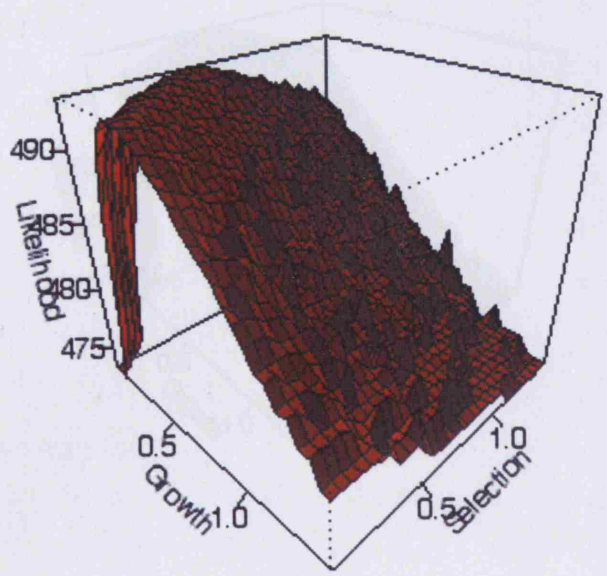
Of all the parameters, mutation rate seemed likely to have the most impact. Given however that microsatellite mutation rate is known to be variable (Weber and Wong 1993, Rolf and Brinkman 1999), there were limited options available for improving the accuracy of the rate used in Syssiphos. The program incorporates a stepwise model and a correction is used to account for differing mutation rates in alleles of different lengths. Mutation rate is known to be affected by the size of the nucleotide repeat unit, (Stumpf and Goldstein 2001) and so for each population group, a series of simulations were run using only the dinucleotide loci and, accordingly, a mutation rate specific for dinucleotides (0.00056) was used (Weber and Wong 1993). For the other simulations, an averaged mutation rate of 0.0012 (Weber and Wong 1993) was used. These data are discussed and shown on table 6.11 in the next section.

6.3.2.1 Changes in run numbers

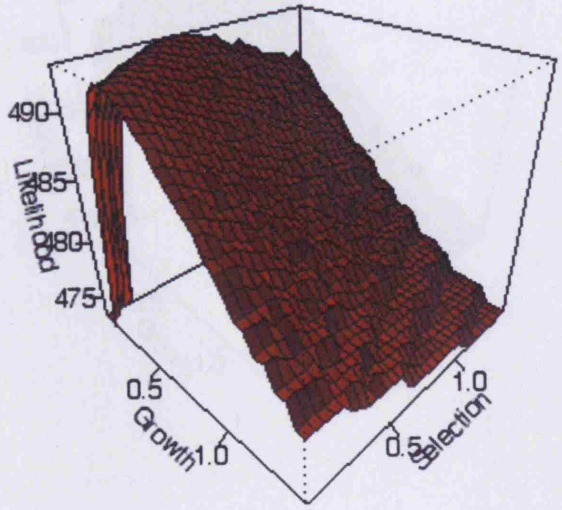
The Syssiphos program can be implemented with shorter or longer run numbers Higher run numbers increase the resolution of the output, but also the time taken to run the program[40]. This can be seen very clearly in figures 6.9 – 6.12, which show the three-dimensional contour maps for each simulation. Simulations for the same data set (Western Europe, ATC haplotype) were run for 10, 100, 1000 and 10,000 times. In each case, the higher simulation number did produce a higher resolution

[40] Depending on the size of the data set, times varied from approximately 2 minutes for a 10 run simulation, over 12 hours for 10,000 runs

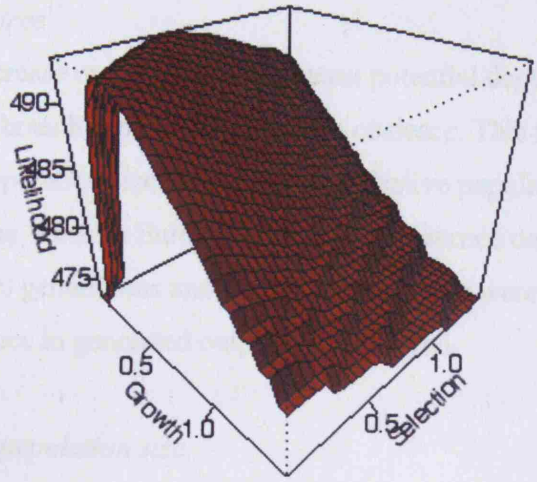
output, which can be seen as a smoother graph (fig. 6.12). To create these figures, the code written for R described in 2.6 and 6.2 was adjusted increase the cut off point to 20 likelihood units of the maximum, to enable a larger upper threshold curve to be visible in the graphical output, so that the differences between the run simulations can be seen more clearly. The data are presented here as three dimensional plots, and, as the figures show, increasing the run number smooths out the distribution of likelihoods on the graph. For the initial estimates of growth and selection for all the data sets, a run number of 100 was used to provide a quick rough estimate. Then, simulations of interest were repeated with a higher run number of 10,000.



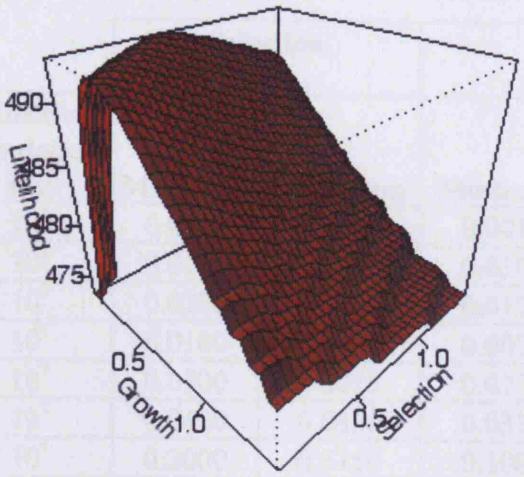
[10 runs]



[100 runs]



[1000 runs]



[10,000 runs]

Figures 6.9 –6.12 A series of three-dimensional plots of a simulation run in Syssiphos for Western Europe, using the ATC SNP haplotype. *The graphs show maximum likelihood value plotted against selection and growth*

6.3.2.3 Depth of tree

It is possible to increase or decrease the greatest potential depth of tree, that is, the point at which the branches of a simulated tree coalesce. This itself is also limited by the effective population size. Assuming an effective population size of 10^7, and using, as before, the Western European data set, coalesence dates of 10,000 generations, 50,000 generations and 100,000 generations were simulated. No significant difference in generated output was observed.

6.3.2.4 Effective population size

The analyses of Caldwell (2005) who used Syssiphos to investigate selection acting on the amylase gene complex, suggests that changing effective population size does not make a significant difference to the Syssiphos output. For lactase however, some substantial differences in the output generated were seen. Again, using the same input data set and keeping all other parameters constant, effective population sizes of 10^4, 10^7 and 10^9 were compared in a series of simulations.

		Selection		Growth	
SNP Haplotype	**Effective population size**	**Minimum**	**Maximum**	**Minimum**	**Maximum**
GCC	10^4	0.0010	0.0175	0.0010	0.0125
GCC	10^7	0.0150	0.0400	0.0100	0.0325
GCC	10^9	0.0200	0.0525	0.0175	0.0500
GCT	10^4	0.0100	0.0300	0.0075	0.0225
GCT	10^7	0.0300	0.0615	0.0275	0.0550
GCT	10^9	0.0400	0.0850	0.0375	0.0750
ATC	10^4	0.2000	0.5750	0.1000	0.3750
ATC	10^7	0.4000	1.1500	0.2500	0.6000
ATC	10^9	0.5000	1.4000	0.3750	0.8000

Table 6.10 - A table to show the selection and population growth minimums and maximums for a series of simulations assuming different effective population sizes. Three different haplotypes in Western Europe were all investigated, as shown above.

Summary of experimenting with the Syssiphos program

It appears that, as might be expected, changing the assumed effective population size has a major impact on the likelihood values generated. The maximum coalescence time of the simulated trees in the program depends partly on Ne, and this is reflected in the program outputs. However, although there are clear differences between the likelihood values generated, suggesting that the selection and growth values themselves are sensitive to the assumed effective population size, these differences seem to affect each SNP haplotype simulation proportionately. Similarly, altering the number of runs that Syssiphos performs in order to estimate the likelihood values for a data set also affects output. As can be seen from the graphs 6.9-6.12, a smoother distribution of values is observed as run numbers increase. Unfortunately, increasing the run numbers also significantly increases the length of time of each simulation, so it was not always practical to perform longer runs. Interestingly, the maximum depth of the tree specified did not significantly alter the likelihood estimates, possibly because all the values supplied were, to be conservative, well beyond the actual coalescence time of the simulated trees.

6.3.3 Testing the hypothesis that selection has acted on the Lactase gene region

A set of simulations was undertaken in which mutation rate and microsatellite loci used was varied for each SNP haplotype.

Syssiphos was operated with the run number set at 100 to give a rough, low resolution estimate of the maximum likelihoods the data for various combinations of selection and population growth. The first tranche of these experiments used all five microsatellite loci as summarised in 6.2 and on table 6.11, and an averaged mutation rate of 0.0012 (Weber and Wong 1993). The second excluded the D2S2385 microsatellite because of its distance from the lactase gene and because the data shown in figs. 6.3-6.8 indicates that recombination may have eroded its linkage-disqeuilibrium with the other microsatellite alleles and the studied SNPs in and around the lactase gene. Another set of simulations excluded both the D2S2385 data and also the tetranucleotide microsatellite, MSAT2, so that a more accurate

mutation rate specific to dinucleotides 0.00056 could be used (Weber and Wong 1993).

The results of these initial experiments are summarised in table 6.11. The table shows, for each simulation for each grouped population, the range of selection values (those producing a likelihood of the data within 2 units of the maximum) assuming population growth is zero, and, the range of population growth values (those producing a likelihood of the data within 2 units of the maximum) assuming selection is zero. Comparing the results for each core SNP haplotype, a signature of possible selection was, as described earlier, found when the minimum likely selection observed for one SNP-defined haplotype was greater than the maximum for the other SNP-defined haplotype(s) in a given population group.

From these simulations, it is clear that the ATC SNP haplotype values are markedly different from the other two in the Eurasian, Western European and the world groups. From inspection of these data, it would appear that the Eurasian selection signature is due to the contribution of the Western European data. Although changing mutation rate and loci used altered the values for selection and growth, selection was consistently observed in these groups and for this haplotype. Selection was not observed for the ATC haplotype in the Ashkenazi population group, or for African data set, which included the Algerian group with just a few ATC chromosomes. In populations where the ATC haplotype was not found, the GCC and GCT haplotypes were compared, but neither fulfilled the criteria for selection.

On the basis of these results, a series of simulations were repeated with a greater number of runs, increasing the accuracy of the likelihood estimates.

Description of simulation				Selection		Growth	
Population	SNP Haplotype	No. of MSAT loci	Mutation rate used	Minimum	Maximum	Minimum	Maximum
Ashkenazi	GCC	3	0.00056	0.0025	0.0100	0.0025	0.0100
Ashkenazi	GCT	3	0.00056	0.0025	0.0150	0.0025	0.0135
Ashkenazi	ATC	3	0.00056	0.0010	0.0175	0.0010	0.0150
Ashkenazi	GCC	4	0.00120	0.0175	0.0475	0.0150	0.0400
Ashkenazi	GCT	4	0.00120	0.0200	0.0038	0.0175	0.0325
Ashkenazi	ATC	4	0.00120	0.0100	0.0425	0.0075	0.0400
Ashkenazi	GCC	5	0.00120	0.0190	0.0440	0.0150	0.0385
Ashkenazi	GCT	5	0.00120	0.0150	0.3100	0.0125	0.0265
Ashkenazi	ATC	5	0.00120	0.0147	0.0490	0.0150	0.0425
Ethiopia	GCC	3	0.00056	0.0010	0.0100	0.0010	0.0100
Ethiopia	GCT	3	0.00056	0.0010	0.0075	0.0010	0.0075
Ethiopia	GCC	4	0.00120	0.0100	0.0220	0.0100	0.0200
Ethiopia	GCT	4	0.00120	0.0100	0.0215	0.0100	0.0185
Ethiopia	GCC	5	0.00120	0.0075	0.0250	0.0075	0.0200
Ethiopia	GCT	5	0.00120	0.0125	0.0200	0.0100	0.0200
Eurasia	GCC	3	0.00056	0.0050	0.0150	0.0050	0.0115
Eurasia	GCT	3	0.00056	0.0175	0.0250	0.0125	0.0225
Eurasia	**ATC**	**3**	**0.00056**	**0.0750**	**0.2750**	**0.0750**	**0.2250**
Eurasia	GCC	4	0.00120	0.0200	0.0395	0.0175	0.0300
Eurasia	GCT	4	0.00120	0.0375	0.0600	0.0300	0.0500
Eurasia	**ATC**	**4**	**0.00120**	**0.3750**	**1.0500**	**0.2500**	**0.6500**
Eurasia	GCC	5	0.00120	0.0200	0.0375	0.0195	0.0325
Eurasia	GCT	5	0.00120	0.0325	0.0500	0.0300	0.0425
Eurasia	**ATC**	**5**	**0.00120**	**0.1000**	**0.1525**	**0.0850**	**0.1850**
Middle East	GCC	3	0.00056	0.0050	0.0185	0.0050	0.0750
Middle East	GCT	3	0.00056	0.0100	0.0200	0.0100	0.0175
Middle East	GCC	4	0.00120	0.0200	0.0520	0.0200	0.0500
Middle East	GCT	4	0.00120	0.0250	0.0415	0.0215	0.0350
Middle East	GCC	5	0.00120	0.0200	0.0500	0.0200	0.0425
Middle East	GCT	5	0.00120	0.0250	0.0450	0.0225	0.0350
Northern Africa	GCC	3	0.00056	0.0001	0.0050	0.0001	0.0085
Northern Africa	GCT	3	0.00056	0.0001	0.0050	0.0001	0.0075
Northern Africa	GCC	4	0.00120	0.0150	0.0225	0.0100	0.0175
Northern Africa	GCT	4	0.00120	0.0100	0.0225	0.0075	0.0200
NorthernAfrica	GCC	5	0.00120	0.0075	0.0225	0.0050	0.0200
Northern Africa	GCT	5	0.00120	0.0125	0.0225	0.0100	0.0175
Western Europe	GCC	3	0.00056	0.0050	0.0125	0.0025	0.0125
Western Europe	GCT	3	0.00056	0.0015	0.0150	0.0050	0.0100
Western Europe	**ATC**	**3**	**0.00056**	**0.1000**	**0.3000**	**0.0750**	**0.2250**
Western Europe	GCC	4	0.00120	0.0175	0.0390	0.0150	0.0350

Description of simulation				Selection		Growth	
Population	SNP Haplotype	No. of MSAT loci	Mutation rate	Minimum	Maximum	Minimum	Maximum
Western Europe	GCT	4	0.00120	0.0300	0.0620	0.0250	0.0550
Western Europe	**ATC**	**4**	**0.00120**	**0.4000**	**1.0500**	**0.2750**	**0.6000**
Western Europe	GCC	5	0.00120	0.0150	0.0350	0.0125	0.0325
Western Europe	GCT	5	0.00120	0.0300	0.0595	0.0255	0.0300
Western Europe	**ATC**	**5**	**0.00120**	**0.1150**	**0.1550**	**0.0850**	**0.1400**
Whole data-set	GCC	3	0.00056	0.0025	0.0100	0.0025	0.0100
Whole data-set	GCT	3	0.00056	0.0100	0.0195	0.0100	0.0150
Whole data-set	**ATC**	**3**	**0.00056**	**0.1000**	**0.2250**	**0.0900**	**0.1750**
Whole data-set	GCC	4	0.00120	0.0175	0.0300	0.0150	0.0250
Whole data-set	GCT	4	0.00120	0.0300	0.0400	0.0250	0.0350
Whole data-set	**ATC**	**4**	**0.00120**	**0.3000**	**0.8000**	**0.2100**	**0.5750**
Whole data-set	GCC	5	0.00120	0.0010	0.0020	0.0010	0.0020
Whole data-set	GCT	5	0.00120	0.0275	0.0400	0.0225	0.0325
Whole data-set	**ATC**	**5**	**0.00120**	**0.1000**	**0.1750**	**0.0750**	**0.1450**

6.11 A table to show the results of the various combinations of Syssiphos simulations. *The results suggesting a positive selection signature are shown in bold. Mutation rate is shown as per (Weber and Wong 1993) per locus per generation. Two different values, 0.0012 or 0.00056 are used depending on whether the microsatellite loci include the tetranucleotide or just dinucleotide loci as shown on the table.*

6.3.4 *Fine-tune investigation of selection*

The final tranche of experiments involved zooming in on a range of values of selection and population growth and increasing the number of runs (10,000) to obtain greater accuracy in the likelihood estimation. D2S2385 was excluded from the simulations, and the averaged mutation rate of 0.0012 (Weber and Wong 1993) was used.

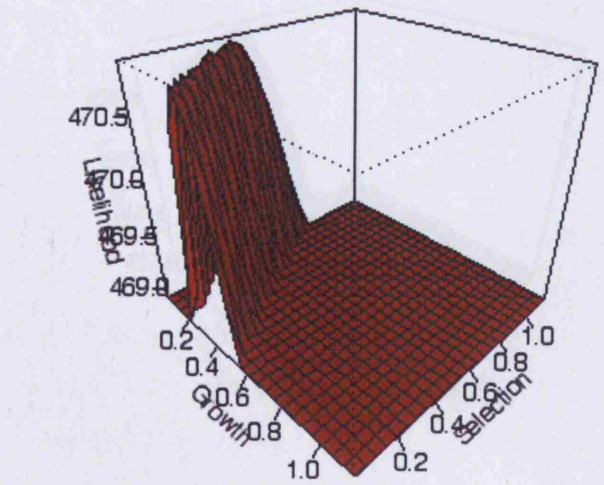
		Selection		Growth	
Population group	**SNP haplotype**	**Minimum**	**Maximum**	**Minimum**	**Maximum**
Western Europe	GCC	0.0125	**0.0400**	0.0125	0.0315
Western Europe	GCT	0.0275	**0.0600**	0.0225	0.0500
Western Europe	ATC	**0.4000**	1.1500	0.2500	0.6000
Middle East	GCC	0.0215	0.0600	0.0150	0.0500
Middle East	GCT	0.0275	0.0425	0.0200	0.0350
Ashkenazi	GCC	0.0275	0.0800	0.0250	0.0675
Ashkenazi	GCT	0.0400	0.0800	0.0350	0.0675
Ashkenazi	ATC	0.0120	0.0400	0.0100	0.0375
Ethiopia	GCC	0.0200	0.0350	0.0175	0.0275
Ethiopia	GCT	0.0125	0.0200	0.0120	0.0200
World	GCC	0.0200	**0.0350**	0.0175	0.0315
World	GCT	0.0250	**0.0400**	0.0250	0.0350
World	ATC	**0.3000**	0.8000	0.2150	0.5500

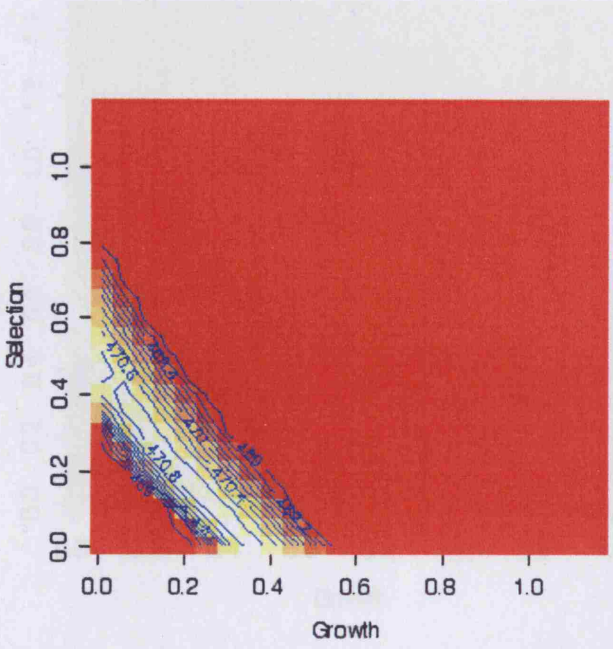
Table 6.12 A table to show the maximum likelihood values of maximum and minimum selection and growth values for a series of population groups calculated at a higher resolution. *The figures in bold show, for selection, that the minimum value for the condensed ATC haplotype is higher than the maximum value for the GCC and GCT haplotpes in the same population group.*

Table 6.12 shows a summary of the range of values of selection or population growth that produce likelihoods of the data not less than 2 units less than the maximum likelihood. As with the initial simulations, these analyses indicate a role for selection in determining the frequency of the ATC haplotype in the Western European group, and, as a result of this, for this same chromosome in the world

group as a whole. Figs 6.13 – 6.16 show the two and three dimensional output for the ATC haplotype in both the World data set, and the Western European data set. The crest of the outputs in figs 6.13-6.16 show the peak of likelihood, which can be seen on the graph as a particular combination of selection and growth. On the two-dimensionl graphs, the whitest region of the image is the area of highest likelihood combinations of selection and growth. As before, it should be noted that because population growth and selection are confounded, near maximum likelihoods are found for a range of value combinations for these parameters. This gives the impression of a 'ridge' of near-maximum likelihoods running diagonally across the graphs.

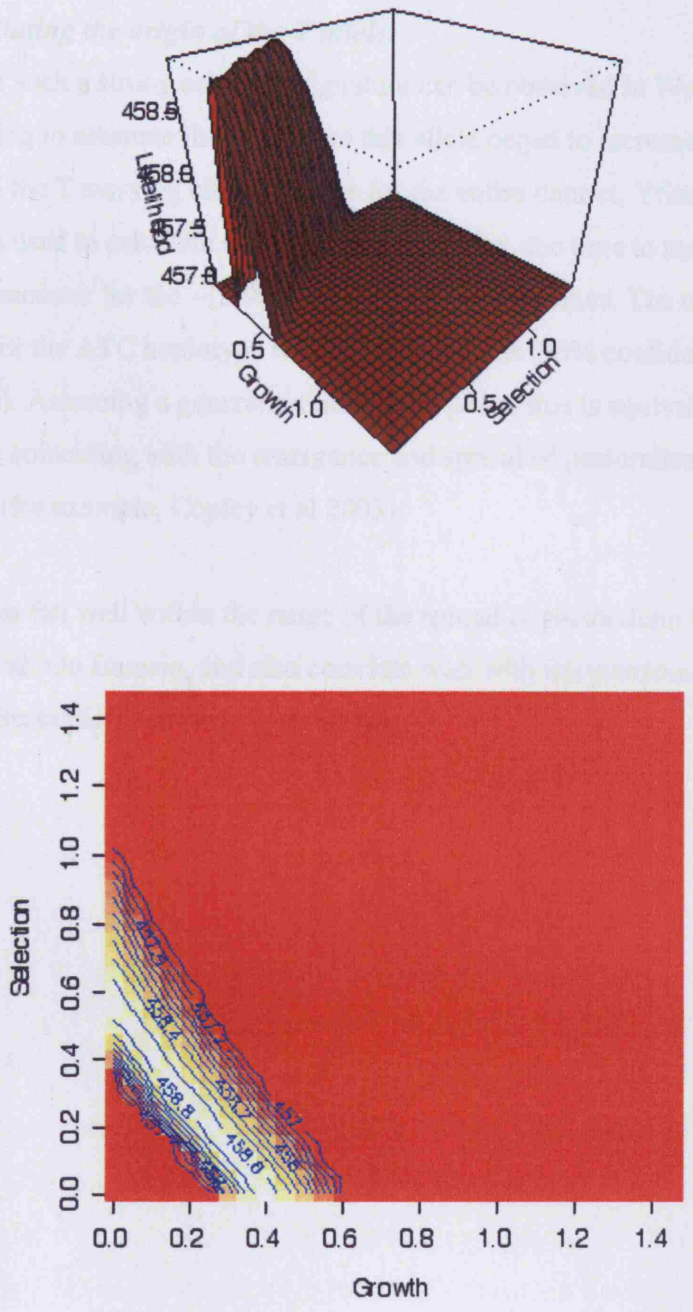
Complete data set – all ATC chromosomes





*Figs 6.13 and 6.14 Figures to show the 3-D and 2-D graphical output of the selection signature for the –13.9kb*T carrying chromosomes in the complete data set*

Western Europe – ATC chromosomes



Figs 6.15 and 6.16 Figures to show the 3-D and 2-D graphical output of the selection signature for the –13.9kb*T carrying chromosomes in the Western European group

6.3.5 Dating the origin of the T allele.

Given that such a strong selection signature can be observed in Western Europe, it is interesting to estimate the time since this allele began to increase in frequency. Taking all the T carrying chromosomes for the entire dataset, Ytime (Behar et al., 2003) was used to calculate the ASD, and from that, the time to most recent common ancestor for the –13.9kb*T carrying chromosomes. The unbiased T_{MRCA} estimate for the ATC haplotype was **351** generations (95% confidence intervals, 299 – 410). Assuming a generation time of 25 years, this is equivalent to 8770 years ago, coinciding with the emergence and spread of pastoralism during the Neolithic (for example, Copley et al 2003).

These dates fall well within the range of the spread of pastoralism out of the fertile crescent and into Eurasia, and also correlate well with the previously estimated date given by Bersaglieri and colleagues (2004).

6.4 Discussion

The key conclusion of this chapter is that it supplies strong evidence of a signature of selection acting on –13.9kb*T carrying chromosomes in Western Europe. This supports the conclusions of Bersaglieri and colleagues (2004), using a different (but related) methodology, based on the accumulation of microsatellite variability rather than the accumulation of recombination events.

Examination of the raw microsatellite data showed that there were some unusual patterns. Looking at microsatellite variation within each of the three SNP haplotypes, it is clear that there was far less diversity for the ATC haplotype than for the GCC and GCT haplotypes. When the distant microsatellite marker D2S2385 was removed, this trend was more marked; for dating, and further investigation, simulations without this microsatellite were used on the basis recombination had disassociated D2S2385 alleles from other loci. The next chapter addresses the issue of where linkage disequilibrium breaks down upstream and downstream of the lactase gene. The data from this chapter strongly suggest that D2S2385 is beyond the main LD block surrounding the lactase gene.

Additionally, although the AMOVA showed significantly more microsatellite variation within population groups than between them, as was expected from studies of other loci, this difference was substantially less than observed previously for these same population groups. Caldwell (2005), using the same family sets[41] used here, showed that 1.07% of variation in microsatellites closely linked to the amylase gene cluster was aportioned between population groups and 98.9% within populations. This result corresponds better with previous research discussed in the introduction (Rosenberg et al 2002), which suggested a ratio of difference in the region of 95% to 5% difference. Similarly, R_{ST} and the Exact Test of Population Differentiation showed significant differences between the population groups,

[41] Because this study and Caldwell (2005)'s study had a small number of dropped samples due to PCR failures, the individuals are not identical although the population groups as a whole are the same.

suggesting that the history of polymorphic sites in and around the lactase gene are atypical.

The maximum likelihood output from the Syssiphos program shows that chromosomes carrying the –13.9kb*T allele exist at a very high frequency in relation to their accumulated microsatellite diversity, a proxy for haplotype age, as indirectly determined by a comparatively low inter-allelic diversity. This strongly supports the hypothesis of historic selection acting on these chromosomes. This selection signature also appears when looking at the complete data set of ATC chromosomes sampled from across the world.

By analysing the Western European data set, the selection signature was apparent in the models even when the parameters of mutation rate, effective population size, run number and microsatellite loci were changed. However, these factors did affect the selection coefficient obtained, the exact value of this is uncertain. It is possible that this sensitivity is only noticeable in cases where there is real and strong selection, since Caldwell (2005) did not observe the same trend when investigating variation in the amylase gene cluster where selection was not detected. The lowest possible range of selection values observed, assuming no population growth was 0.2000 – 0.5750, when effective population size was set at 10^4, mutation rate at 0.0012 and 4 microsatellite loci were used. The highest was observed as 0.5000 – 1.4000, when effective population size was 10^9 assuming population growth is zero.

The selection minimum and maximums reported in table 6.11 and 6.12 both assume no growth, so that a comparison between different estimates of selection for the different SNP haplotypes can be made. In actuality, the value of population growth will be higher than 0, and so selection will be lower than the minimum and maximum values will be lower than those shown on table 6.12. Figures 6.14 and 6.16 show different ranges of selection given different values of growth. Most significantly, it should be assumed that population growth is equal for each of the

three SNP haplotypes, so the comparative values of selection can be determined based on a constant growth value.

When the Ashkenazi Jewish and the Algerian –13.9kb*T carrying chromosomes were analysed independently of the dataset for the rest of Europe, no evidence for selection was found. However, the modal microsatellite haplotype associated with the ATC chromosomes in Europe, **6, 7, 15, 11**[42] (taking four loci) was found in both these groups (8/8 in the Ashkenazis, 4/6 in the Algerians). The most likely explanation for this failure to uncover evidence for selection in this restricted data set is that the sample sizes, given the number of –13.9kb*T alleles, were too low. However, it may also be that the -13.9kb*T carrying chromosomes in Algerian and amongst the Ashkenazi Jewish groups have not been under selection. Their occurrence in these population groups may be explained by demographic factors, particularly admixture. It is known that the Ashkenazi Jews in particular have a complex history of migration, and admixture with non-Jewish populations may have introduced these alleles into the population (Thomas et al, 2002).

Since lactase persistence chromosomes might occur on the background of an A haplotype chromosome predating the emergence and spread of the –13.9kb*T allele, the marker most strongly associated with the A Haplotype, 5579*C, was also investigated for evidence of selection. Neither the Ethiopian nor the Middle Eastern samples, where the –13.9kb*T allele was not observed, showed evidence of selection for the 5579*C. However, it is not known is how many lactase persistent chromosomes were present in these population groups, and also the data from the previous chapter suggests in the Middle East especially, lactase persistence may occur on a different haplotypic background. As discussed in the previous chapter, further exploration of polymorphism associating with lactase persistence outside of Western Europe may reveal different alleles, and it would be interesting to see whether Syssiphos could identify a selection signature when such loci are considered.

[42] The four loci are: msat2, intron1, msat 3 and msat4 as described in table 6.1 and figure 6.1

If, however, genetic drift shaped by demographic history were responsible for the distribution of lactase polymorphisms, it might be expected that the effect would be seen at other loci. The Syssiphos program has also been used on the same samples[43] to look for evidence of selection, but for two different regions of the genome; the amylase gene cluster (Caldwell 2005) and the CYP2D6 allelic frequency (Stumpf and Fletcher, personal communication). However, comparable data on these regions showed no evidence of selection. The observations described here for the *LCT* and *MCM6* gene region are consistent with the findings of Bersaglieri et al (2004), who showed outlier F_{ST} status for the G-22kbA and C-13.9kbT polymorphisms when compared to 28,400 other markers located throughout the genome.

The Syssiphos program does not model episodic major historical bottlenecks, though work is currently underway to incorporate such demographic events into the program. It is thought that the most recent bottlenecks will have the strongest effects on a population in terms of program modelling (Michel Stumpf, personal communication).

Looking at the distribution of microsatellite haplotypes in the population groups, and for each of the SNP haplotypes, it is clear that in each population group there is one prevalent microsatellite haplotype. Specifically, it can be seen that, for the ancestral SNP haplotype of GCT, the common microsatellite haplotype **7, 10, 17, 11** was found at relatively high frequency in the Armenians (0.33) the Ashkenazis (0.30) and the Ethiopians (0.26). From studying the microsatellite data in the French CEPHS and the Finnish population, where core SNP lactase haplotypes were known, it is possible to identify this as being a signature microsatellite haplotype of the B SNP haplotype.

[43] Although the same DNA samples were used, Caldwell (2005) and Fletcher (2002) had cases of different individual samples failing, so the numbers and data are not identical for the population groups as a whole.

As a whole, from this data set, it appears that the GCC SNP haplotype in the Amharic Ethiopians, Ashkenazi Jews and Middle Eastern groups did not show evidence of being under selection. The 5579*C allele defines the lactase core A Haplotype in Europe, and a possibility was considered that, if –13.9kb*T allele were not causative of lactase persistence, the 5579*C might associate with lactase persistence in populations where –13.9kb*T was absent. However, from this data set, there was no evidence from the Syssiphos simulations of selection acting on 5579*C alleles.

Bersaglieri and colleagues provided estimates of age of –13.9kb*T carrying chromosomes, with a broad range of possible dates, which were also used to determine the selection coefficient. Using a microsatellite-based technique, dating the T_{MRCA} for the ATC SNP haplotype in Europe gave a credible figure of approximately 351 generations (95% confidence intervals, 299 – 410) which would also place the emergence of these –13.9kb*T carrying chromosomes during the Neolithic. This supports the hypothesis that the selection event in Western Europe occurred following the introduction of dairying and fresh milk drinking (which may or may not have coincided with the spread of pastoralism) into the region from the Near East.

Subsequent to the work in this chapter being done, a paper by Coehlo et al (2005) reported similar data and a similar method to investigate selection pressures in a series of populations. The authors genotyped both the G–22kbA and C-13.9kbT SNPs and four microsatellite polymorphisms in a series of unrelated individuals from Portugal, Italy and three Fulbe groups from Mozambique, São Tomé and Cameroon. 794 chromosomes were resolved using PHASE (Stephens et al 2001) Using a method based on intra-allelic diversity (Slatkin and Bertorelle 2001) and under the same assumptions as described in this chapter, they investigated whether the observed frequency of the -22kbA and -13.9kbT alleles are consistent (given an expectation of neutrality) with intra-allelic diversity as measured by the four linked microsatellite loci.

Haplotype networks were generated by the NETWORK program v.4.0.0.0[44] in order to estimate the minimum number of mutations (S_O) between the linked loci necessary to produce the observed haplotypes. Estimates of allelic age were made for each population using two methods: the first, as in this chapter, used Ytime to date the T_{MRCA}, using two mutation rates. In addition, Coelho and colleagues also used another method based on the decrease in frequency of the modal allele, which enabled them to take into account recombination, and again, two different mutation rates were used (Coelho et al., 2005).

These estimates of T_{MRCA} for the –13.9kb*T allele were used in conjuntion with two different demographic models to test whether frequency and diversity departed from expectations under neutrality. The first demographic model was that of a constant growth rate of population, at a rate of 10^3, over 900 generations. The second, also over 900 generations, used an exponential growth rate of 10^4 to $5x10^9$, which might be more consistent with population expansions during the Neolithic (Coehlo et al 2005). Using T_{MRCA} based on both high and low mutation rates, and both demographic models, Coehlo et al (2005) were able to reject neutrality in for the Portugese and the Fulbe ($p => 0.001$). However, in the Fulbe of São Tomé, neutrality could only be rejected in the models where the higher mutation rate was used. In the Italian group, neutrality could not be rejected when the first demographic model (that of a constant growth rate) was considered with the lower mutation rate, as this combination reduced the expected intra-allelic diversity.

Looking at all the possible age estimates for the pooled data set, the authors concluded that the –13.9kb*T allele was likely to have emerged prior to the Neolithic, but after the expansion of modern humans out of Africa. Using Ytime, ages between 45,000 and 30,000 years ago were made using the higher mutation rate, and 17,500-11,750 for the lower. The alternative methodology suggested a T_{MRCA} of between 12,300 – 7,450 years.

[44] This program is available from http://www.fluxus-engineering.com

One possible reason for the descrepency between the T_{MRCA} estimates of Coehlo et al (2005) and those presented here is that they used PHASE software (Stephens et al 2001) resolve haplotypes. If this software should fail to resolve haplotypes correctly then it is likely that it would overestimate the microsatellite diversity associated with the -13.9kb*T allele. This would have the effect of pushing back (overestimating) T_{MRCA} for the -13.9kb*T carrying chromosomes. Nonetheless, consistent with the analysis presented here, they concluded that there is strong evidence of selection acting on the -13.9kb*T allele in European groups. Interestingly, they also identified evidence for selection acting on the -13.9kb*T allele in the Fulbe. The further implications discussed in the paper will be considered in the final chapter.

Chapter Seven

Linkage disequilibrium across the lactase gene

7.1 Introduction

Recent evidence, particularly from Bersaglieri et al (2004), has shown significantly longer shared haplotypes for chromosomes carrying the –13.9kb*T allele than those without. For many of these, haplotype blocks of approximately 1MB have been observed (Poulter et al 2003, Bersaglieri et al 2004). Comparison with other regions of the genome suggests this may represent a departure from the usual situation. For example, although a genome-wide study suggested that 50% of the genome exists in blocks these are mostly tens of kb in size rather than many hundreds (Gabriel et al 2002).

The region around lactase also has a comparatively low recombination fraction of (0.2/0.3)[45]. If the –13.9kb*T allele is not causal of lactase persistence, it could be tightly linked with an as yet unknown mutation on the background of an extended A Haplotype, itself within an extensive region of linkage disequilibrium (LD).

The study by Bersaglieri and colleagues (2004) suggested that it is only the chromosomes carrying the –13.9kb*T alleles that have unusually extended haplotypes, but the non –13.9kb*T alleles were not subdivided further with respect of the *LCT* SNP core haplotypes. The evidence from Poulter et al (2003) suggested that at least in the French CEPH sample, some extended -13.9*C A haplotype chromosomes and possibly some extended non A haplotype chromosomes might be present. This chapter uses data made publicly available by the HapMap project to investigate how far LD extends in relation to the core lactase haplotypes in the Utah CEPH samples. The chapter also examines association between core lactase haplotypes and the two *MCM6* alleles described by Enattah et al (2002) with a distant marker in different population groups.

[45] Sex-averaged genethon and DeCode data from the Human Genome Browser, that is, assuming 0.5cM per Mb as downloaded from the July 2003 freeze.

7.2 HapMap data

The International HapMap project aims to categorise genetic data from, to date, four populations: unrelated Han Chinese individuals from Beijing, unrelated Japanese individuals from Tokyo, a series of Utah CEPH families and a series of Yoruba families from Ibadan in Nigeria. Genotype information from these groups can be used to compare genetic variation in different groups of modern humans, with the ultimate objective of providing a tool for researchers interested in mapping genes related to disease. Data downloaded from the HapMap website[46], was used as a resource to investigate and characterise the extent of linkage disequilibrium around the lactase gene, and to identify areas of interest for future study.

7.2.1 Linkage disequilibrium in the Utah CEPHs

A high-density map of SNP polymorphisms[47] was used to investigate LD between polymorphic sites in and around the lactase gene. In total, data from a series of 494 SNPs characterising a region of 2MB were initially downloaded for the 30 trios taken from the Utah CEPH families (see fig 4.1). This included the lactase gene, position: 136,756,613-136,801,327 (5' to 3')[48] on chromosome 2, and surrounding areas upstream and downstream, located between positions 135,780,851 and 137,780,850. Data produced by researchers in the Galton laboratory on markers in the *LCT* gene was already available for more than 2/3 of these families (Harvey 1994, Hollox 2000).

[46] The website is: http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap
[47] The SNPs downloaded were approximately 5kb apart
[48] The positions given are those cited on the HapMap website, determined using the Genome Browser from the July 2003 freeze of this data.

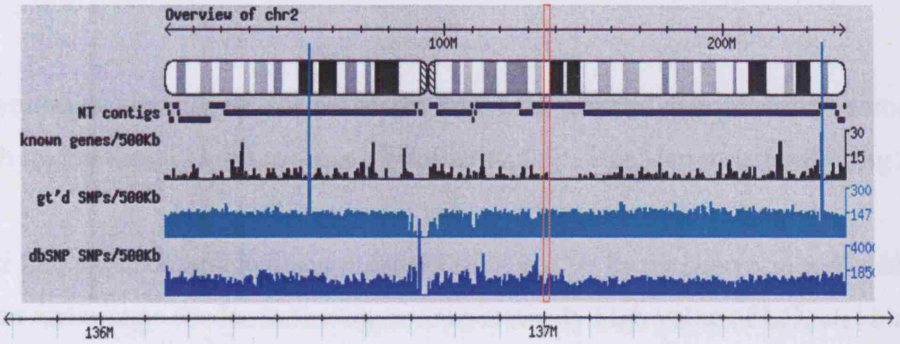


Fig 7.1 A diagram to show a 2MB region of interest including and surrounding the lactase gene

The red bar shown at 137M indicates the section of chromosome 2 from which a series of SNPs were downloaded from the HapMap website

Polymorphic sites that showed allele frequencies of less than 7% for this sample were excluded. The data for the remaining 346 polymorphisms were tabulated and entered into the Quickstart programme[49] to generate an input file for PHAMILY[50] which was used to determine phase. Extended haplotypes for the 120 parental chromosomes were generated and were named 'HapMap haplotypes' to differentiate them from other haplotypes described in this thesis.

Once the parental chromosomal haplotypes had been established, they were used as input data for the HaploXT program[51] (Abecasis and Cookson 2000), which can be used to determine linkage disequilibrium where phase is known. The linkage disequilibrium for the region was then shown using the programme Gold[52] (Abecasis and Cookson 2000), which produces a visual representation of a specified measure of LD[53], taking into account the physical location of each

---

[49] This programme is available at the following website: http://archimedes.well.ox.ac.uk/pise/quickstart.html

[50] This programme is available at the following website: http://archimedes.well.ox.ac.uk/cgi-bin/pise/lib/connect.pl

[51] Found at website: http://archimedes.well.ox.ac.uk/cgi-bin/pise/haploxt.pl

[52] Found online at website: http://www.sph.umich.edu/csg/abecasis/GOLD/index.html. Gold can also display severeal LD statistics, such as *chi* square and associated p-values, r^2, D, D', Cramer and U.

polymorphic site. Figure 7.2 shows the pattern of linkage disequilibrium across the 2Mb region for all chromosomes downloaded from HapMap measured using r^2.

Loci that are in complete disequilibrium (that is, $r^2 = 1$) are shown in red, with green and orange shades indicating a comparatively high value of LD, and blue regions showing low LD. The lactase gene itself is located at position 136,756,613-136,801,327 (5' to 3'), and *MCM6* at 136,807,966-136,844,779 (5' to 3'). Figure 7.2, it can be clearly seen that there is a very large region of LD, covering more than 1Mb but within that block there is some breakdown of LD such that there are two smaller blocks, one of which shows higher values and more areas of complete LD.

The smaller block contains *LCT*, *MCM6* and surrounding markers and covers about 300-400kb while, surprisingly the larger block with the higher values of LD is downstream from *LCT*. The distant marker, D2S2385, which from the data in chapter 6 appeared not to be in strong LD with *LCT*, is located about 200kb further upstream than the region shown (located approximately at position 137,9726,830), well beyond the large region of LD visible on this map.

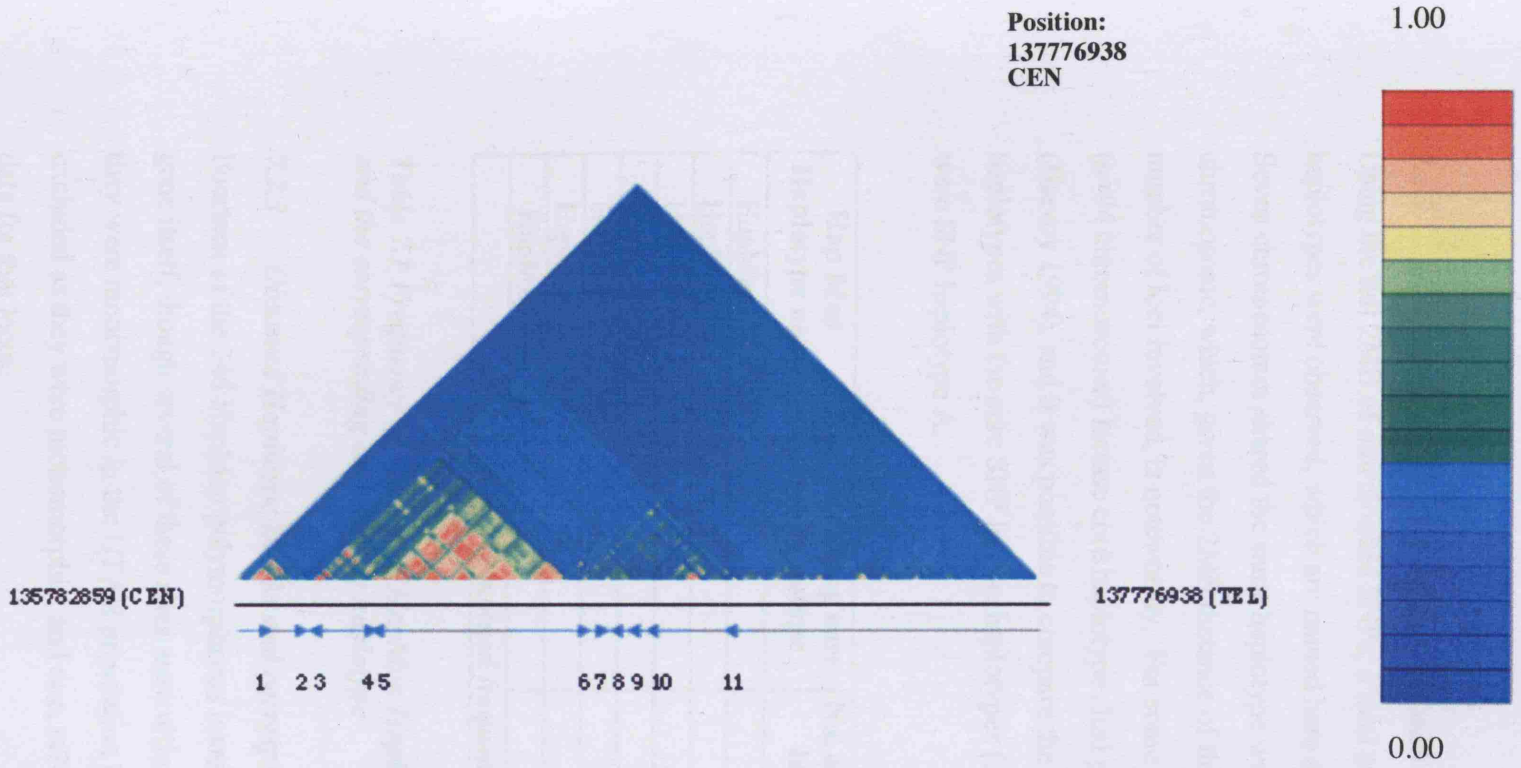


Fig 7.2 r^2 values across a 2MB region around the lactase gene

The scale to the right of the LD map shows values of $\mathbf{r^2}$ *against colour, where red or '1.00' indicates complete LD.*
TEL indicates the telomeric end of the chromosome and CEN the centromeric end.
The blue arrows at the base of the diagram represent genes and the direction of their transcription, and the number below each arrow corresponds to the name of the gene as follows:
1 = ACMSD, 2 = CCNT, 3 = FLJ23074, 4 = RAB3GAT, 5 = ZRANB3, 6 = R3HCM, 7 = UXBD2, 8 = LCT, 9 = MCM6, 10 = DARS and 11 = CXCR4

7.2.2 *Association of HapMap haplotypes with core lactase haplotypes*

Using the full 2MB of downloaded SNPs, a total of 103 different HapMap haplotypes were observed, which are named here as HapMap 1 to HapMap 103. Seven chromosomes shared the same haplotype were found in more than one chromosome, which, given the 2MB distance of the region under analysis and the number of loci involved, is noteworthy. For some of the Utah CEPH individuals, (n=94 chromosomes) lactase core haplotypes had previously been characterised (Harvey 1994), and it was possible to compare the more frequent HapMap haplotypes with the core SNP lactase haplotypes (see table 7.1). In all cases they were SNP haplotype A.

Hap Map Haplotype name	Corresponding core lactase haplotype	No. of Chromosomes in population	Frequency (n = 120)
HapMap 20	A	8	0.067
HapMap 29	A	5	0.042
HapMap 41	A	3	0.025
HapMap 67	A	2	0.017
HapMap 43	A	2	0.017
HapMap 37	A	2	0.017
HapMap 4	A	2	0.017
Total combined frequency: 0.2			

Table 7.1 Frequency of a series of HapMap Haplotypes and the corresponding core lactase haplotype

7.2.3 *Observed Haplotype Blocks and corresponding core lactase haplotypes*

Fourteen of the 346 HapMap polymorphisms investigated are located in the lactase gene itself, though several of these sites were ultimately excluded in this study as they were monomorphic in the UTAH population (see fig 7.3). Of these, five were excluded as they were monomorphic, and one, rs745500, because of insufficient data for that locus.

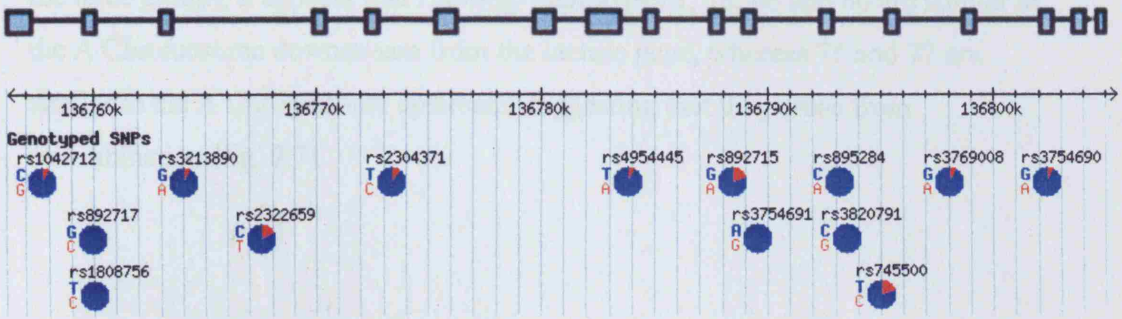


Fig 7.3 – Distribution and frequency of HapMap alleles located in the lactase gene adapted from the HapMap web-site

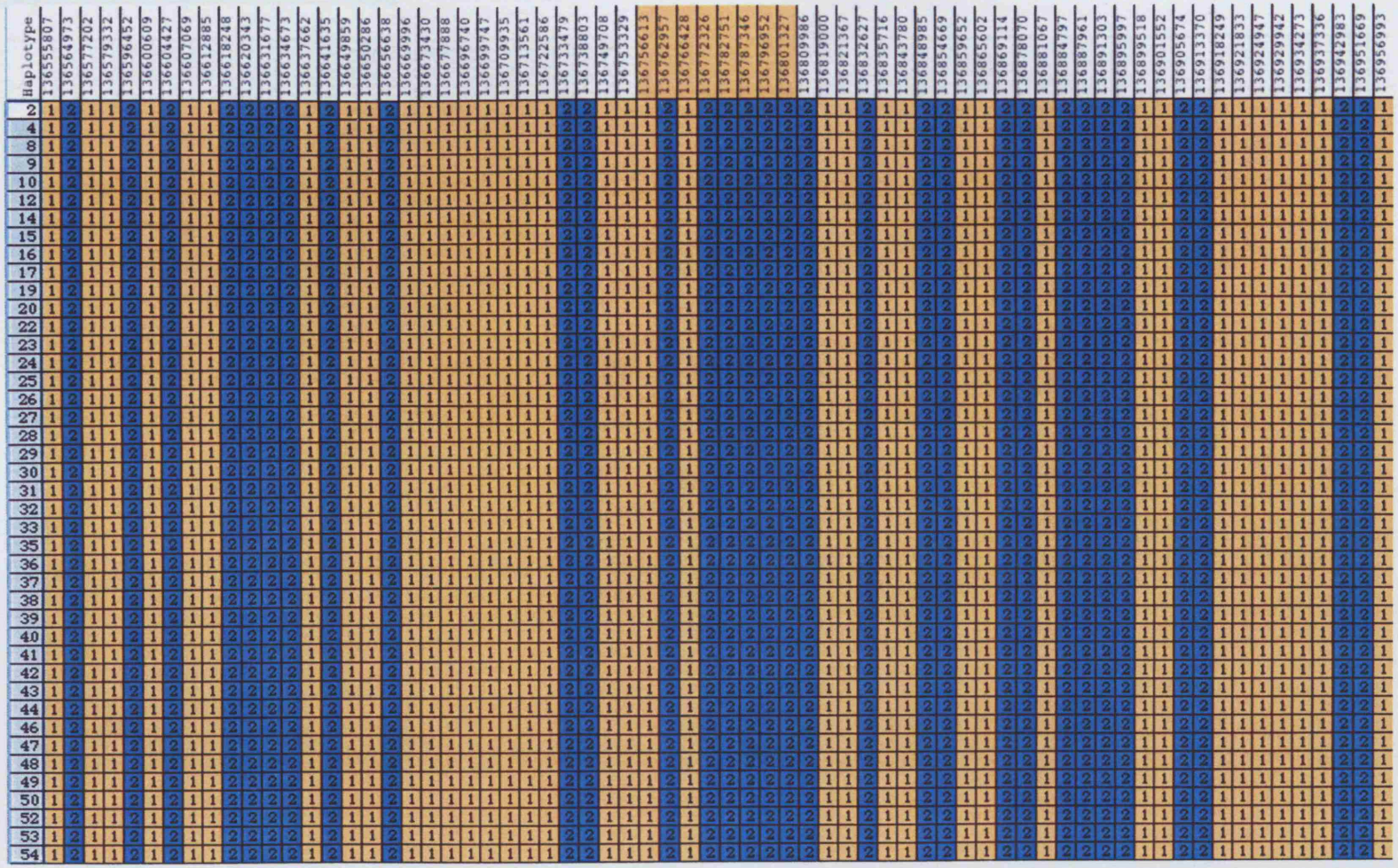
Polymorphic markers are those showing diversity in allele frequencies, orientated against a schematic image of the lactase gene above, with exons indicated by blue squares and introns by lines.

Using core lactase haplotype data (Harvey 1994), and the HapMap loci specifically located in the lactase gene itself, it was possible to correlate the A, B and C haplotypes with those generated by HapMap. HapMap chromosomes were first grouped by known core lactase haplotype, (A, B or C) and, within these groups, had identical allelic states for those markers within the lactase gene itself.

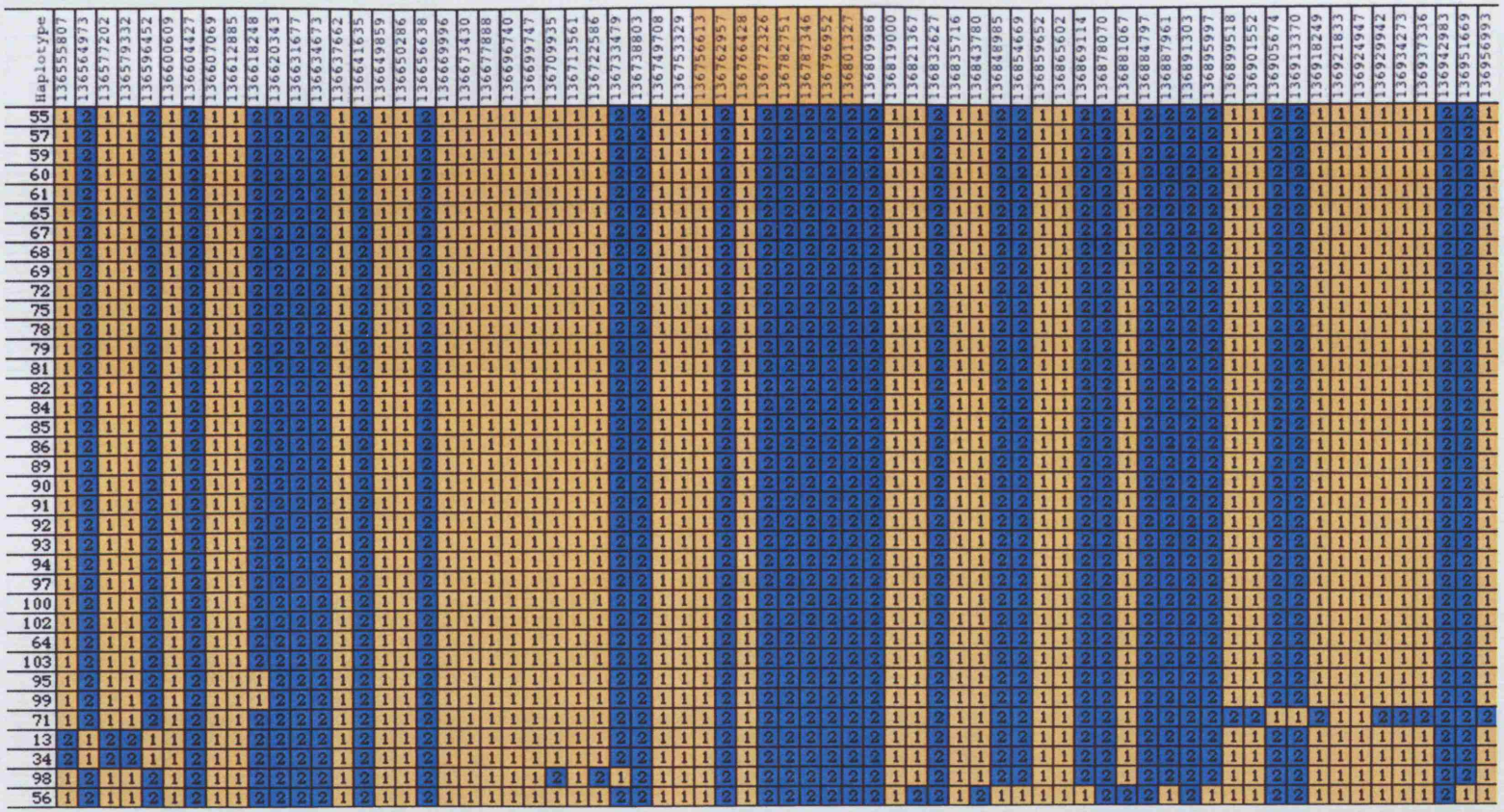
To increase the sample size, chromosomes for which the core lactase haplotype was unknown were sorted into group A, B or C if they had identical allelic states for those markers located in the lactase gene. An extended chart showing 30 polymorphic sites either side of the lactase gene (corresponding to 400kb) was created, with the HapMap haplotype grouped by the core lactase haplotypes (see figs. 7.4-7.7). These results confirm that, there is far more homozygosity (or there is a far more extended region of haplotype identity) for the A Haplotype than the B and C haplotypes. Only 7/78 of the A chromosomes show evidence of breakdown of haplotype across this region, while half (5/10) of the B chromosomes and 7/9 of the C chromosomes do so.

the three groups, it appears that HapMap haplotypes 1, 18, 66 and 80 are similar to the A Chromosome downstream from the lactase gene, whereas 76 and 77 are similar to the A Chromosome upstream, suggesting that they arose from recombination (fig. 7.7).

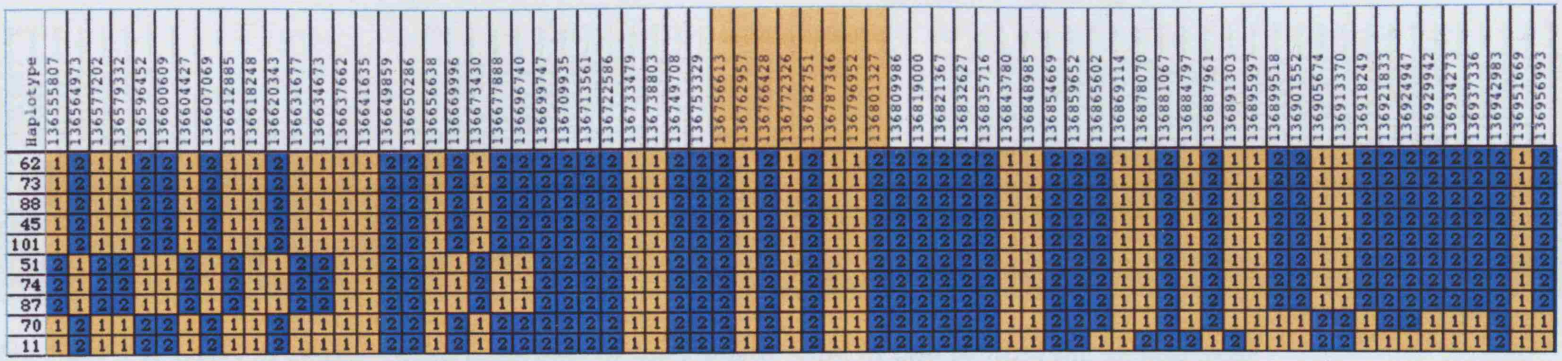
A Haplotype Chromosomes – 7.4(i)



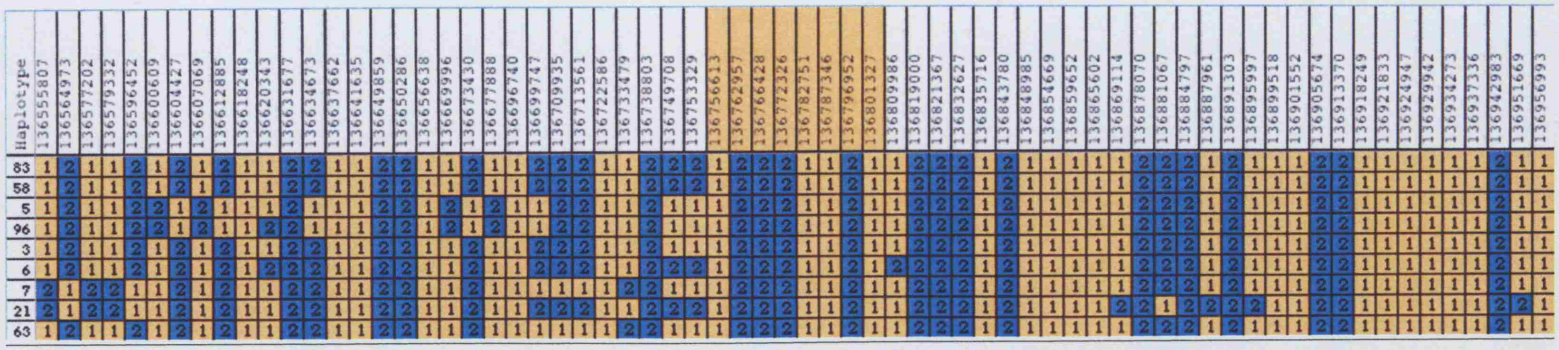
A Haplotype Chromosomes – 7.4(ii)



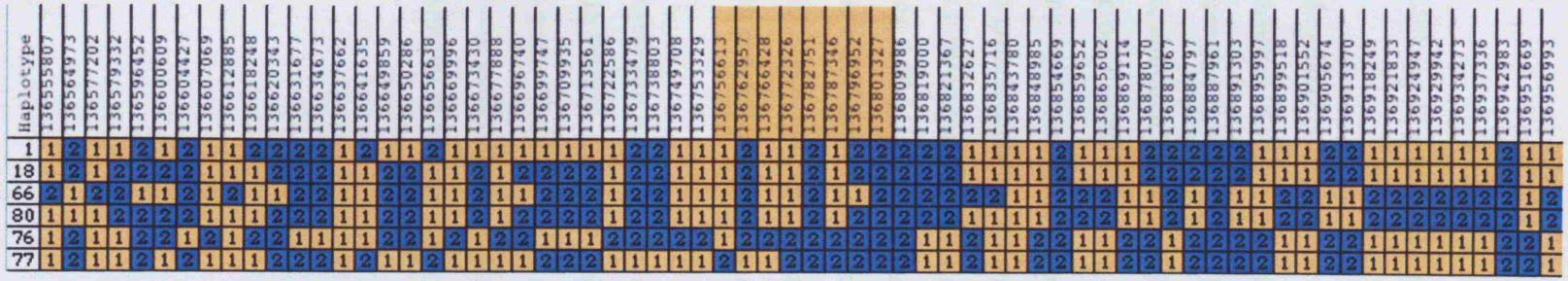
B Haplotype Chromosomes –7.5



C Haplotype Chromosomes -7.6



Recombinant/Unassignable haplotypes – 7.7:



Figs 7.4-7.7 A series of colour-coded diagrams to show allelic state in and around the lactase gene for HapMap chromosomes grouped by LCT core haplotypes.

The positions of the markers are given at the top of the chart, with those inside the lactase gene shaded in yellow. The HapMap chromosome haplotypes are given on the side.

7.3 Linkage disequilibrium measured using a distant marker

The HapMap data confirmed that the very frequent A Haplotype in the Utah CEPH group is identical in allelic state over a very long region (400kb shown in Fig 7.4) in 91% cases and that this is not the situation for the other 2 core lactase haplotypes, although half of the B haplotype chromosomes were identical for markers across this 400kb region. Inspection of pairwise LD values across the lactase gene region, made available by HapMap, shows that these differ between population groups (data not shown). It was therefore of interest to determine whether A Haplotypes carrying the –13.9kb*T allele show a longer range of identity than A Haplotypes carrying the –13.9kb*C allele. It was not possible to address this question with the Utah CEPH sample because information regarding the –13.9kb*T polymorphism was unavailable and there were in any case likely to be very few non-13.9kb*T A chromosomes in this population.

The most upstream marker[54] that defines the core *LCT* extended A haplotype, described in Poulter et al (2003), is an A – G transition, located –370kb upstream of the lactase gene. The location of this marker is indicated in figure 7.8, beyond the most upstream gene of that region, *CXCR4*. This marker was investigated in a series of population samples where core lactase haplotype was known.

[54] This polymorphic site was first described in Poulter et al 2003 as 'Marker 3'

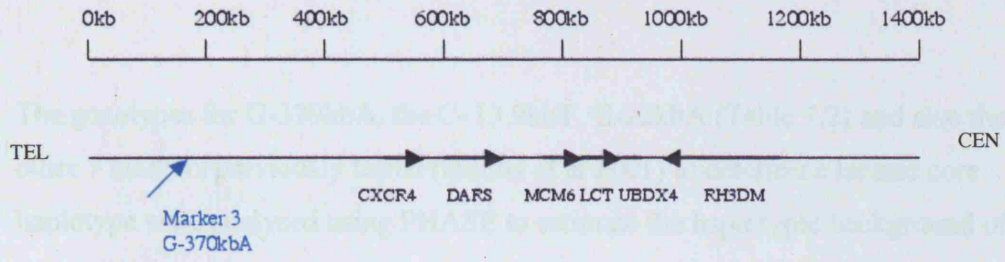


Fig 7.8 A diagram to show the location of the G-370kbA polymorphism located 370kb upstream of the transcriptional start of the lactase gene. *The diagram scale is shown above, and the black arrows represent genes and the direction of their transcription. G-370kbA or Marker 3 is shown in blue, and located in an inter-gene region. Tel and Cen indicate the orientation of the chromosome, 'telomeric' and 'centromeric'*

To determine the association between this marker, –13.9kb*T and the core lactase haplotypes, a series of different populations (the Roma, North and South Indians, Malay, Chinese, Japanese, South African Bantu and San - described in chapter 3) were typed (see fig 7.9 and section 2.4).

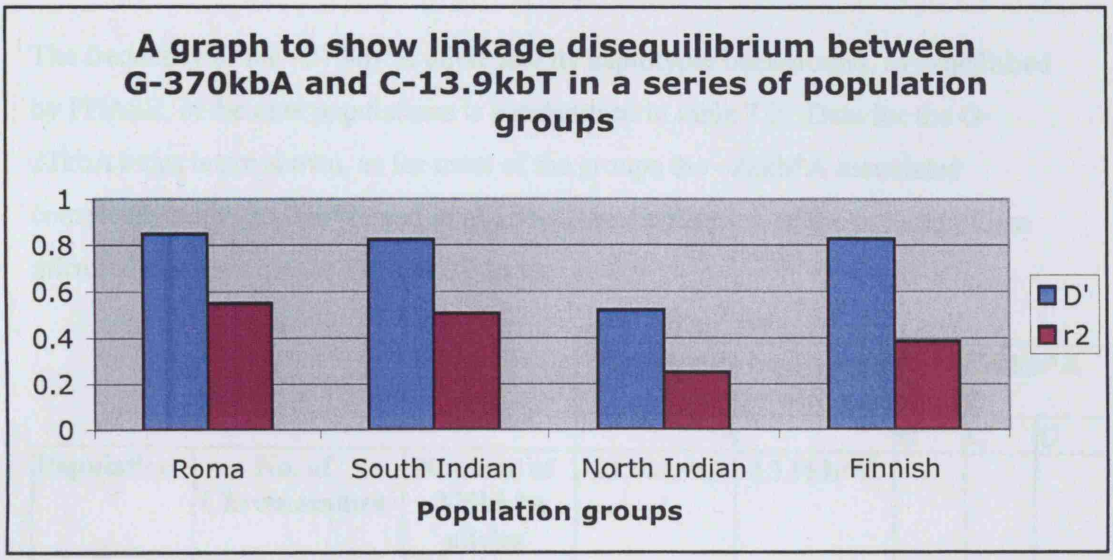
Figure 7.9– An example gel for typing the G-370kbA polymorphism.
Lane 1 shows a ladder of fragments of known DNA size. Lane 2 is the negative control, lanes 3 and 4 show a gel band of 324bp, indicating a gel phenotype of a homozygote (G) allele, and lane 5 shows a gel phenotype of a heterozygote, with the additional (A) allele inferred by the two smaller bands of 180bp and 144bp

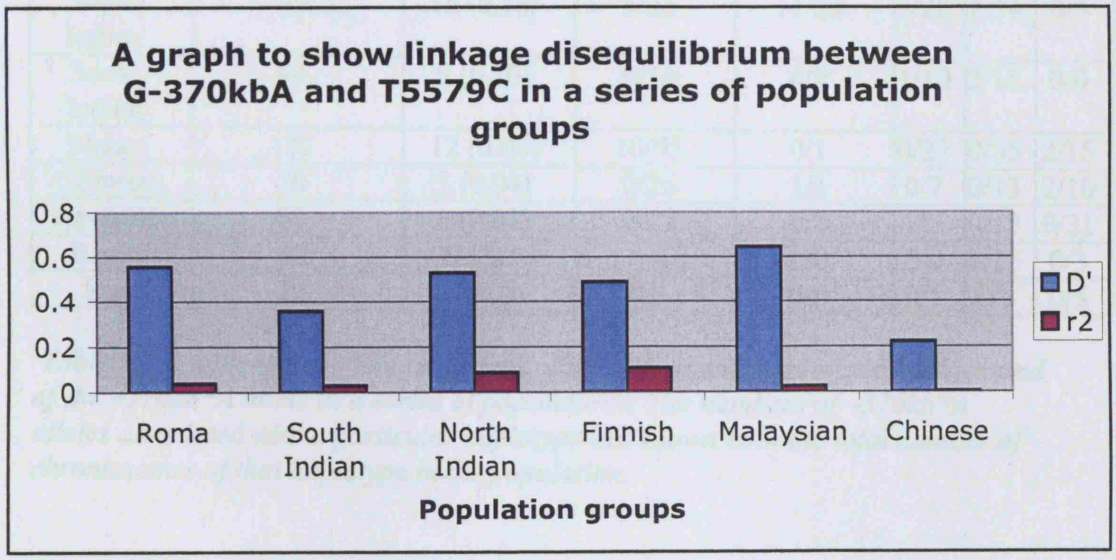
The genotypes for G-370kbA, the C–13.9kbT, G-22kbA (Table 7.2) and also the other 7 markers previously tested (Hollox et al 2001) to determine lactase core haplotype were analysed using PHASE to estimate the haplotypic background of the -370kb*A polymorphism. HaploXT was then used to establish linkage disequilibrium.

Figures 7.10 and 7.11 show the linkage disequilibrium, measured in both D' and r^2, for pairwise comparisons between G-370kbA and both the C-13.9kbT and the T5579C loci. The San and Bantu were monomorphic for both G-370kbA and the C-13.9kbT, hence no data is shown, and similarly, some of the Southeast Asian populations showed extremely low frequency of the derived alleles, and so LD results are omitted.

A pair wise comparison was undertaken for all four Eurasian groups where both C-13.9kbT and G-380kbA were polymorphic. C-13.9kbT and G-370kbA showed statistically significant association (based on χ^2 associated p-values, ($p = < 0.05$). D' was moderately high (< 0.5) for C-13.9kbT and G-370kb. The r^2 values suggest fairly strong LD among Roma and South Indian groups but a relatively weaker association with values of less than 0.5 among the Finnish and North Indian groups.

For the T5579kbC allele and the G-370kbA, association as determined by χ^2 was shown to be statistically significant in the Roma, North Indian, Finnish and Malaysian groups ($p = < 0.05$). In general, values of D' were high, whereas r^2 values, all of less than 0.5, indicated a somewhat weaker association.





Figs. 7.10 and 7.11 Two graphs illustrating linkage disequilibrium between loci for a series of population groups

The blue bars show D' values, and the red bars show r^2 values. In the top graph, all values are also significant p=<0.05. In the lower graph, all but the South Indian and the Chinese are significant p =<0.05

The frequency of the –370kb*A allele and its haplotypic background, as established by PHASE, in the nine populations is summarised in table 7.2. Data for the G-22kbA locus is not shown, as for most of the groups the –22kb*A associated completely with –13.9kb*T, and in all cases, the distribution of the polymorphism mirrored that seen for the C-13.9kbT locus.

			Haplotypic background of –370kb*A allele where observed				
			A		B	C	U
Population	**No. of Chromosomes**	**Number of –370kb*A alleles (frequency)**	**-13.9kb*C**	**-13.9kb*T**			
Finnish	72	28 (0.39)	5/22	17/19	0/17	6/13	0/0
Roma	154	19 (0.12)	2/61	13/16	4/47	0/16	0/0
North Indian	102	18 (0.18)	2/26	11/24	3/28	2/34	0/0
South Indian	44	9 (0.20)	0/14	6/9	1/10	2/16	0/0
Malay	192	12 (0.06)	10/95	0/1	0/27	0/35	2/15
Chinese	70	3 (0.04)	0/26	1/1	0/7	0/13	2/10
Japanese	80	1 (0.01)	0/29	0/0	1/5	0/13	0/21
Bantu	40	0 (0.0)	0/6	0/0	0/0	0/15	0/3
San	28	0 (0.0)	0/2	0/0	0/1	0/1	0/5

Table 7.2 A table to show the frequency, distribution and haplotypic background of the –370kb*A allele in a series of populations. *The numbers of –370kb*A alleles associated with a particular haplotype are shown with the total number of chromosomes of that haplotype in the population.*

The –370kb*A allele occurs mainly, though not exclusively, on an A Haplotype background, and on some B, C and U haplotypes. Considering only A Haplotype chromosomes, it is more frequently seen with –13.9kb*T allele, but it is appears as though there has been some historic recombination between the two loci, since all four combinations of alleles (that is, all four combinations of –13.9kb*C/T and –370kb*A/G) are found even within a single population, such as the Finns.

7.4 Linkage Disequilibrium in Finns of known lactase persistence status

The 5 microsatellite loci, and 3 SNP loci used in chapter 6 (see also 6.2 and 2.4) were typed in a series of Finnish individuals (see also 3.2). This Finnish sample, for which lactase persistence status was known, was divided into persistent (n = 32) and non-persistent (n = 40)[55]. PHASE was used as before to establish haplotypic background, then HaploXT to determine linkage disequilibrium in an attempt to see whether these parameters differed between the two groups.

The LD scores for each pair wise comparison are summarised in figures 7.12 – 7.13 In the non-persistent group, the C-13.9kbT and G-22kbA loci were, as indicated, monomorphic. Perhaps surprisingly given the findings of chapter 6, $\chi 2$ comparisons showed significant LD for the D2S2385 marker and other polymorphic loci: in the non-persistent group, D' was significant between D2S2385 and intron 1, Msat3 and Msat4, and in the persistent group. Considering the p-values generated by HaploXT, there were a higher number of significant observations for LD in the persistent group, most markedly for Msat3, and also the T5579C locus defining the A Haplotype.

[55] It should be noted that the numbers of persistant and non-persistant individuals reported here cannot be used to represent actual lactose persistance frequencies in the Finnish population, since the group was self-selected by individuals believing themselves to be lactose malabsorbers.

Non-Persistent Group

	D2S2385							
MSAT2	0.345	MSAT2						
G-22kbA	/	/	G-22kbA					
C-13.9kbT	/	/	/	C-13.9kbT				
Intron1	**0.542**	**0.772**	/	/	Intron1			
MSAT3	**0.470**	**0.731**	/	/	**0.905**	MSAT3		
MSAT4	**1.000**	1.000	/	/	1.000	1.000	MSAT4	
T5579C	0.312	**0.599**	/	/	**0.792**	**0.750**	1.000	T5579C

Persistent Group

	D2S2385							
MSAT2	**0.857**	MSAT2						
G-22kbA	0.478	0.439	G-22kbA					
C-13.9kbT	0.478	0.439	**1.000**	C-13.9kbT				
Intron1	0.540	0.515	0.495	0.495	Intron1			
MSAT3	**0.635**	**0.833**	**0.832**	**0.832**	**0.644**	MSAT3		
MSAT4	0.828	1.000	0.439	0.439	0.210	**0.667**	MSAT4	
T5579C	0.526	**0.590**	**0.719**	**0.719**	**0.641**	**0.923**	1.000	T5579C

Fig 7.12 and Fig 7.13 Two tables to show the D' values for a series of pair wise comparisons between loci in two groups of Finns, persistent (n = 32) and non=persistent (n = 40) *Values shown in bold showed a significance from a generated HaploXT p value (p = < 0.05)*
'/' indicates values were excluded as the polymorphism was monomorphic in this group.

7.5 Discussion

The data from this chapter confirms the observation of extended LD in and around the lactase gene in certain haplotypes for certain populations. The Utah CEPH data suggests that the extended A haplotype (of >1MB), first observed in Poulter et al (2003), is uniquely conserved when compared to the other core lactase haplotypes commonly found in Europe (B and C). As a whole, the Utah CEPH data when investigated for LD across a 2MB region suggests that there are distinct haplotype blocks punctuated by 'hot spot' regions, in keeping with observations about the pattern of LD in other regions of the genome (for example, Goldstein et al 2003). It is also the case, though, that the Utah CEPHs are a comparatively homogenous population, and results may also have been influenced by low allelic frequencies.

Where the –13.9kb*T allele was observed, it appeared to be in LD with the –370kb*A allele in the Roma, North and South Indian groups and Finnish population, as might be expected. This suggests that A Haplotypes carrying the –13.9kb*T allele probably do show a longer region of shared allelic identity than those without, in these populations as well. Where –13.9kb*T allele was not found, it appears as though there are some differences between population groups for LD with the –370kb*A associating with the 5579kb*C in some groups but not others. This observation may suggest that the molecular history of the A Haplotype differs between population groups. To investigate this possibility, groups known to have high lactase persistence frequencies but which do not carry the alleles associated with the trait in Europe could be targeted for more extensive LD tests using a wider array of markers.

In the Finnish data set there were little evidence of significant difference in LD between the persistent and non-persistent groups but there were some features of interest in the pair wise LD analyses. One was the very distant microsatellite marker D2S2385, since there seemed to be more significant associations between it and other markers in the non-persistent group. This association might reflect the

presence of unusually long haplotypes in this population, but this is not consistent with the comparatively weak association with marker 3.

The markers used in this thesis, and in particular when investigating selection, have been centred in the lactase gene or regions upstream of it. The choice of markers was, in part, made on the expectation of the causality of the –13.9kb*T allele, or its close association with a causal mutation, possibly in the same vicinity. However, the unusual findings of a large block of linkage disequilibrium downstream of the lactase gene indicates that it would also be interesting to investigate more distant markers, further downstream of the lactase gene, to understand more about this region.

There are several possible explanations for the observation that LD is higher downstream of the lactase gene: first, the Utah CEPHs are, as discussed, a relatively homogenous population group, and the high LD downstream of lactase may be influenced by this low intra-population diversity. Another possibility is of suppression of recombination, preserving LD downstream of the lactase gene. Further possibilities, which are more extreme, are that a causal mutation for lactase persistence is located on the other side of the gene. However, this would contradict the interpretations of other current research (for example, Bersaglieri et al 2004).

Chapter Eight

Discussion and Conclusions

The core aim of this thesis was to investigate whether the cultural shift documented as the 'Neolithic revolution' and associated changes in human demography are reflected in the modern day distribution of genetic variation in and around the lactase gene. During the course of this thesis work, several papers relevant to the evolution of the lactase persistent trait were published (Enattah et al 2002, Bersaglieri et al 2004, Coehlo et al 2005), and these influenced the specific aims ofthe research.

The description of the two polymorphisms located –13.9kb and –22kb upstream of the lactase gene, within introns of the *MCM6* gene (Enattah et al 2002), precipitated further investigation into their relationship with other known lactase polymorphisms. As chapter 3 describes, both the derived –13.9kb*T and –22kb*A alleles were found to associate with the 5579*C allele, which defines the Eurasian A Haplotype described by Harvey et al (1998) and Hollox et al (2001). This pattern of association was consistently found to be the case wherever the –22kb*A and –13.9kb*T alleles were observed. The only exception was found in one Algerian family, where one GTT[56] chromosome was observed, which probably arose due to historic recombination events. Many chromosomes around the world that carry the 5579*C allele, (A Haplotype) do not carry the –22kb*A and –13.9kb*T alleles, indicating that both are relatively recent mutations. An Insertion/Deletion polymorphism located in intron 1 of the lactase gene and first described in Poulter et al (2003) was similarly investigated, and found to sub-divide the A Haplotype, but all –13.9kb*T alleles were associated with the short form of the InDel-intron1 polymorphism.

The most likely genealogical relationships between these alleles associated with lactase persistence is summarised in Figure 8.1.

[56] The GTT condensed haplotype is defined in 3.4, as –22kb*G, -13.9kb*T and 5579*T.

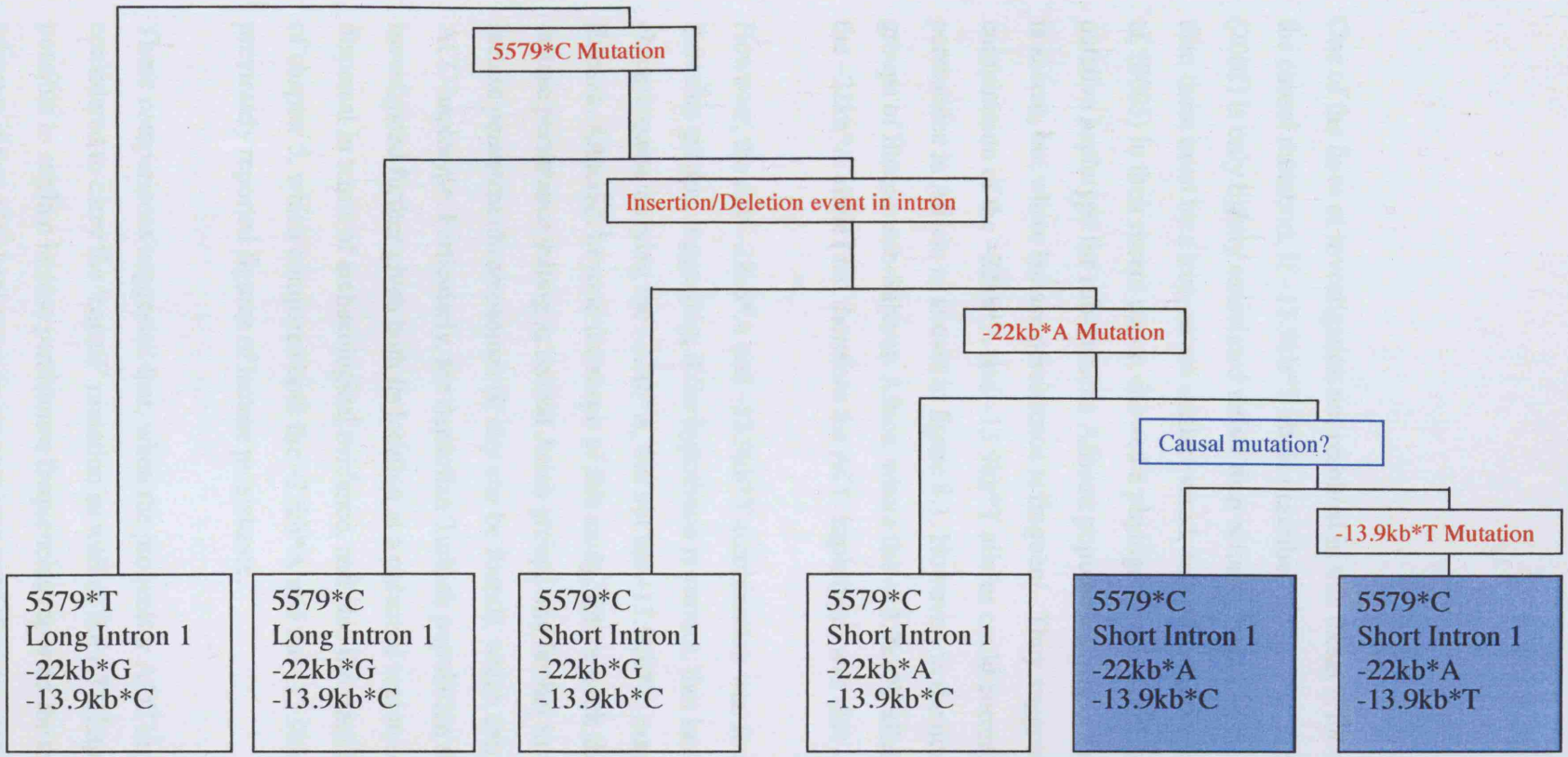


Fig.8.1 A diagram to show the evolutionary relationship between the alleles associated with lactase persistence.
*Mutational events for each polymorphism are shown in red text. Poulter et al (2003) and Coehlo et al (2005) suggest a putative causal mutation, discussed in chapter 5, is shown on the diagram in blue text. The blue shaded haplotypes would be, under this model, both lactase persistent. However, it should be noted that, if there is more than one causal mutation for the lactase perstence trait, an independent causal mutation could have occurred at any branch of this diagram. It is also possible that the –13.9kb*T allele is truly causal, and this would negate the probability of a ACC chromosome carrying the lactase persistent trait.*

One of the lines of investigation not resolved in this thesis is the absolute identity of the causal mutation. If –13.9kb*T allele described by Enattah and colleagues (2002) is only tightly associated rather than actually causal of lactase persistence, then there must be a true, as yet undiscovered, causative mutation exists. Coehlo et al, (2005) in their recent paper, showed a phylogeographic plot which showed a putative haplotype for sub-Saharan African populations where the –13.9kb*T allele is absent, but where lactase persistence is frequent. They suggest that the rare combination of the –22kb*A and –13.9kb*T alleles could correspond with lactase persistence in Africa, as shown in figure 8.1. However, in the nomadic pastoralist groups of Eastern sub-Saharan Africa, where the –13.9kb*T allele was not found, the –22kb*A allele (and therefore the ACC haplotype) was also not found.

However, the rare –22kb*A and –13.9kb*T combination was found in some central Eurasian groups, suggesting, if the hypothesis is correct, that lactase persistent chromosomes carrying the –22kb*A, but not the –13.9kb*T, may exist in central Eurasia. Although beyond the scope of this study, future work should involve lactase persistence testing in central Asian groups and further investigation of lactase persistent chromosomes (if any can be found), which carry the condensed ACC haplotype. Particularly, the Anatolian Turkish population should be investigated further given both its location at a potential region of origin of darying dispersal in terms of archaeological evidence, and also the statistical comparisons of chapter 5, which compared both the –22kb*A and the –13.9kb*T alleles with previously reported figures of lactase persistence.

These comparisons suggested that, when the progenitor ACC haplotype is considered to carry the 'causal' mutation as well as the ATC haplotype, it was possible to explain lactase persistence frequencies reported for the Anatolian Turks, whereas if the ATC haplotype on its own was used, this was not the case. Similarly, groups in Pakistan also showed evidence of the ACC haplotype

(Bersaglieri et al 2004) as did certain Afghanistan groups, and again, given the location of these areas, it would be interesting to sample more widely here.

Equally, such further study might show lactase persistence does not associate with the ACC haplotype in these groups, indicating that the –13.9kb*T allele proposed as a causative mutation may well be so in Eurasia (Enattah et al 2002). In any case, given the distribution of allele frequencies in Eurasia, it seems likely that the –13.9kb*T allele occurred on the background of an A Haplotype carrying the –22kb*A allele.

If, however, the –13.9kb*T allele is truly causal in Europe, given its complete absence amongst East African sub-Saharan pastoralists, and also its inability to predict lactase persistence amongst certain Middle Eastern groups, it is probable that there is another causative mutation. Again, the data in this thesis could support several potential situations, and future work might limit some of these alternative possibilities.

This true causal allele (or even alleles) would, presumably, occur in the Bedouin, Wolof, Nilo-Saharan pastoralist groups and other populations where lactase persistence is observed but neither the –22kb*A nor the –13.9kb*T allele were found. Theoretically, this or these independently evolved mutations might even be trans-acting. It is also possible (though unlikely) that lactase persistence caused by this alternative mutational event might not show the same pattern of inheritance seen in Europe, that is, of a cis-acting autosomal dominant trait.

Under this hypothesis, lactase persistent chromosomes would reach an intermediate to high frequency independently in different regions with a history of pastoralism, evolving congruently with European lactase persistence. Such independent evolution of a trait has a precedent in the case of sickle cell anaemia (Pagnier et al 1984, Tishkoff et al 2001) where a selective advantage occurs. It is also a more

likely scenario if there is a particularly mutable sequence region within a DNA element capable of affecting lactase expression.

An unknown alternative causal mutation could have occurred therefore on the background of the A haplotype prior to the emergence of both MCM6 alleles, as this haplotype, associated with lactase persistence in Europe, is also found throughout Africa and the Middle East (Harvey et al 1998, Hollox et al 2001). The comparative LD and allelic identity of the A Haplotype amongst the French CEPHs (Poulter et al 2003) and also the Utah CEPHs suggests a different evolutionary history from the other core lactase haplotypes in Europe. Considering the distant marker, G-370kbA, it appears that although a significant association could be observed between the derived –370kb*A allele and the-13.9kb*T (in those populations where the T allele was found), this association was less clear for 5579*C allele which defines the A Haplotype. In this case, significant association could be shown between the two alleles in the Roma, Finnish, North Indian and Malaysian population groups, but not in the African groups.

The 5579*C allele showed no significant difference in frequency between the Israeli Bedouin, thought to have a high frequency of lactase persistence, and the neighbouring Israeli Arab community who do not. In Africa, no significant differences in frequency were observed between the Nuer and Anuak, both with different farming economies and dependencies on fresh milk drinking.[57] Nor was the A haplotype frequent enough to explain lactase persistence in any of the populations investigated in this thesis. There were, however, significant associations found between the A Haplotype with pastoralism. A significant association was also found with language phylum; the highest proportion of 5579*C chromosomes was found in the Niger-Congo speakers, (55%), with the lowest in the Nilo-Saharan speakers (25%) and an intermediate frequency in the Afro-Asiatic speakers (44%). It may be the case that amongst the Nilo-Saharan

[57] The assumption that the Bedouin and Nuer are fresh milk drinkers whereas their neighbours, the Israeli Arab community and the Anuak are not, comes from the observations of the sample collectors.

pastoralists of Eastern Africa, and pastoralists in the Middle East, a mutation occurred on an entirely different haplotypic background, and that this has been selected for independently. Again, lactase persistence testing amongst Eastern African and Middle Eastern pastoralists may reveal the presence of lactase persistent chromosomes that are not A haplotype. Such evidence would strongly support independent and convergent evolutionary processes in African and in Europe (Bersaglieri et al 2004, Mulcare et al 2004, Coehlo et al 2005).

The hypothesis that selection favoured lactase persistent individuals able to digest fresh milk was tested by comparing intra-allelic variation of a haplotypes associated with lactase persistence and haplotypes that were not. Although alleles associating with lactase persistence in Europe existed at high frequency, they had low microsatellite diversity. A new analysis program, 'Syssiphos', was used to examine the likelihoods of the data when different values of growth were assumed. The analysis showed that, regardless of population growth, selection could be observed for chromosomes carrying the condensed ATC haplotype. This result supports the work of other recent studies have similarly shown this pattern of selection, both using methods based on intra-allelic diversity (Bersaglieri et al 2004, Coehlo et al 2005). In these cases, the paper by Coehlo and colleauges (2005) rejects neutrality, whereas Bersaglieri et al's study (2004) and the results of chapter six both estimate selection strength.

The strength of the Syssiphos program is its ability to model population growth, microsatellite mutation rate length dependence and recombination. Coehlo et al (2005) used two different demographic models to factor in different growth scenarios, and incorporated a dating procedure using different estimates of mutation rate and recombination. Syssiphos is able to incorporate recombination and growth, and detects selection consistently even when microsatellite mutation rates are varied.

All three investigations showed evidence for selection acting on chromosomes carrying the –13.9kb*T allele in Europe. Interestingly, Coehlo and colleagues were also able to observe this in the Fulbe or Fulani. The data from chapter 4 suggests that a historic introgression of Supra-Saharan chromosomes could explain the presence of the allele; the distribution of the –13.9kb*T did not correlate with location within the Extreme Northern Region of Cameroon, but, again, there was association with Fulani ethnic identity. It is possible that the introgression came from Nomadic pastoralists migrating south from the Mahgreb, as Fulani oral tradition suggests, since –13.9kb*T allele was also found in a Berber and an Algerian population, although at somewhat lower frequency.

In summary there is strong evidence to support the theory that lactase persistence conferred a selective advantage in the Neolithic amongst European populations, and perhaps also the Fulani, where the Y chromosome data supports a prehistoric introgression of Eurasian *LCT/MCM6* gene segment. What has yet to be conclusively demonstrated is the occurrence of an independent causal mutation.

References

References

Abbas H and Ahmad M (1983) Persistence of high intestinal lactase activity in Pakistan Hum Genet 64:277-278

Abecasis GR and Cookson WOC (2000) GOLD – Graphical Overview of Linkage Disequilibrium. Bioinformatics 16:182-183

Agresti A (1992) A survey of exact inference for contingency tables Sta Sci 7(1):131-177

Ahmad M and Flatz G (1984) Prevalence of primary adult lactose malabsorption in Pakistan. Hum Hered 34:69-75

Akey JM, Zhang G (2002) Interrogating a high-density SNP Map for signatures of natural selection. Genome Res 12(12):1805-14

Aoki K (1986) A stochastic model of gene-culture coevolution suggested by the "culture historical hypothesis" for the evolution of lactose absorption in humans. Proc Natl Acad. Sci. USA 83:2929-2933

Asp N, Berg A, Dahlqvist et al (1975) Intestinal disaccharidases in Greenland Eskimos. Scand J Gastroenterol. 10:513-519

Alonso S and Armour J A L (1998) MS205 Minisatellite Diversity in Basques:Evidence for a Pre-Neolithic Component. Gen Res 1290-1298

Ammerman AJ and Cavalli-Sforza LL (1984) The Neolithic Transition and the Genetics of Populations in Europe. Princeton, NJ: Princeton University Press.

Anderson E C and Slatkin M (2004) Population-Genetic Basis of Haplotype Blocks in the 5q31 Region. Am J Hum Genet. 74:40-49

Anderson B and Vullo C (1994) Did malaria select for primary adult lactase deficiency? Gut 35:1487-89

Anh N T, Thuc T K and Welsh J D (1977) Lactose malabsorption in adult Vietnamese. Am J Clin Nutr. 30:468-469

Antonowicz I and Lebenthal E (1977) Developmental pattern of small intestinal enterokinase and disaccharidase activities in the human fetus. Gastroent. 72:1299-1303

Arnold J, Diop M, Kodjovi M, and Rozier J (1980) L'intolerance au lactose chez l'adulte au Senegal. Comptes Rend Soc Biol 174:983-992.

Arribas JCD, Herrero AG, Martin-Lomas M, Canada F J, Shouming HE and Withers SG (2000) Differential mechanism-based labeling and unequivocal activity assignment of the two active sites of intestinal lactase/phlorizin hydrolase. Eur J Biochem 267:6996-7005

Arola H, Koivula T, Jokela H, Jauhiainen M, Keyrilainen O, Ahola T, Uusitalo A, and Isokoski M (1988) Comparison of indirect diagnostic methods for hypolactasia. Scand J.Gastroenterol. 23:351-357.

Asp N-G, Berg N-O, Dahlqvist A and Gudmand-Hoyer E (1975) Intestinal Disaccharidases in Greenland Eskimos. Scand J Gastroent. 5:513-519

Auricchio S, Rubino A and Morset G (1965) Intestinal glycosidase activities in the human embryo, fetus and newborn. Pediatrics 35:944-948

Aumassip, G and G Delibras 1982-3 Ages des depots neolithiques du gisement de Ti-n-Hanakaten (Tassili-n-Ajjer, Algerie) Libyca 30-31, 207-11

Bafna V, Gusfield D, Lancia G and Yooseph S (2003) Haplotyping as Perfect Phylogeny: A Direct Approach (2003) J Comp Biol 10:323-340

Bamshad M and Wooding SP (2003) Signatures of natural selection in the human genome. Nature Reviews Genetics 4:99-111

Barbujani G, Sokal RR and Oden NL (1995) Indo-European origins: a computer-simulation test of five hypotheses. Am J Phy Anth 96(2):109-32

Barbujani G, Magagni A, Minch E and Cavalli-Sforza L L (1997) An apportionment of human DNA diversity. Proc Natl Acad Sci 94:4516-4519

Barich et al (1992) Holocene communities of western and central Sahara: a reappraisal. In New Light on the Northeast African past, F.Klees & R. Kuper (eds) 185-204. Africa Praehistoria 5. Koln: Heinrich-Barth Institut.

Barich,B.E. (1987) (ed) Archaeology and environment in the Libyan Sahara. The excavations in the Tadrat Acacus 1978-1983. BAR International series 368 Oxford BAR.

Barraclough G (ed) (1994) The Times Concise Atlas of World History. Times Books, division of HarperCollins publishers, London.

Bayoumi RAL, Flatz SD, Kuhau W, and Flatz G (1982) Beja and Nilotes: nomadic pastoralist groups with opposite distributions of the adult lactase phenotypes. Am.J.Phys.Anthropol. 58:173-178.

Bayoumi RAL, Saha N, Salih AS, Bakkar AE, and Flatz G (1981) Distribution of the lactase phenotypes in the population of the Democratic Republic of the Sudan. Hum Genet 57:279-281.
Bedine MS and Bayless TM (1973) Intolerance of small amounts of lactose by individuals with low lactase levels. Gastroent. 65:735-743

Behar DM, Thomas MG, Skorecki K, Hammer MF, Bulygina E, Rosengarten D, Jones AL, Held K, Moses V, Goldstein D, Bradman N & Weale ME (2003) Multiple origins of Ashkenazi Levites: Y chromosome evidence for both Near Eastern and European ancestries. American Journal of Human Genetics 73: 768-79

Bersaglieri,T., Sabeti,P.C., Patterson,N., Vanderploeg,T., Schaffner,S.F., Drake,J.A., Rhodes,M., Reich,D.E., and Hirschhorn,J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. 74, 1111-1120.

Bertorelle G and Excoffier L (1998) Inferring Admixture Proportions from Molecular Data Mol Biol Evol 15(10):1298-1311

Beja-Pereira A, Luikart G, England PR, Bradley DG, Jann OC, Bertorelle G, Chamberlain AT, Nunes TP, Metodiev S, Ferrand N and Erhardt G (2003) Gene-culture coevolution between cattle milk protein genes and human lactase genes. Nat Genet 35(4) 311-313

Birge SJ, Keutmann HT, Cuatrecasas P and Whedon GD (1967) Osteoporosis, intestinal lactase deficiency and low dietary calcium intake. New Eng J Med 276:445-448

Biscione, R Dynamics of an early South Asian urbanization: the first Period of Shahr-I-Sokhta and its connections with Southern Turkmenia. Chpt 8 p 105-118 South Asian Archaeology ed. Hammond, N. Gerald Duckworth and company ltd. Duckworth.

Blaxter KL (1961) Lactation and the growth of the young. In: SKKow, AT Cowie (eds), Milk: the mammary gland and its secretion. Academic Press, New York: 329-338

Blench R (1999) Why are there so many pastoral groups in Eastern Africa? In Azarya V et al (eds) Pastoralists Under Pressure? Fulbe Societies Confronting Change in West Africa. Leiden, Boston, Koln, Brill.

Bogucki P (1988) Forest farmers and Stockherders: Early agriculture and its consequences in North-Central Europe. Cambridge:Cambridge University Press.

Bolin T.D., Davies, A.E., Seah C.S., Chua K.L., Yong V, Kho K.M., Siak C.L. and Jacob E (1970) Lactose Intolerance in Singapore. Gastroent. 59(1):76-84

Bolin TD and Davies AE (1969) Asian lactose intolerance and its relation to intake of lactose. Nature 222:382-383

Boll W, Wagner P, and Mantei N (1991) Structure of the chromosomal gene and cDNAs coding for lactase-phlorizin hydrolase in humans with adult-type hypolactasia or persistence of lactase. Am.J.Hum.Genet. 48:889-902.

Bose D P and Welsh J D (1973) Lactose malabsorption in Oklahoma Indians. Am J Clin Nutr. 26:1320-1322

Bozzani A, Penagini R, Velio P, Camboni G, Corbellini A, Quatrini M, Conte D and Bianchi P A (1986) Lactose Malabsorption and Intolerance in Italians. Clinical Implications. Dig Dis Sci. 31(2):1313-1316

Brand J C, Gracey R M , Spargo et al (1983) Lactose malabsorption in Austrailian Aborigines. Am J Clin Nutr 37:449-452

Brinkmann B, Klintschar M, Neuhuber F, Huhne J and Rolf B (1998) Mutation Rate in Human Microsatellites:Influence of the Structure and Length of the Tandem Repeat. Am J Hum Genet. 62:1408-1415

Büller,H.A., Kothe,M.J.C., Goldman,D.A., Grubman,S.A., Sasak,W.V., Matsudaira,P.T., Montgomery,R.K., and Grand,R.J. (1990). Coordinate expression of lactase-phlorizin hydrolase mRNA and enzyme levels in rat intestine during development. J. Biol. Chem. 265, 6978-6983.

Buning, C, Jurga, J, Fiedler, T, Kupferling, S, Worm, M, Weltrich, R, Genschel, J, Lochs, H, Schmidt, H, and Ockenga, J. Genetic Background of Lactose Intolerance and Implications for Diagnosis. Gastroenterology 124 Suppl 1, A-144. 2003. Ref Type: Abstract

Burnham, P (1996) The politics of cultural difference in Northern Cameroon. International African Library. Edinburgh University Press.

Burgio G R, Flatz G, Barbera C, Patane R, Boner A, Cajozzo C and Flatz S D (1984) Prevalence of primary adult lactose malabsorption and awareness of milk intolerance in Italy. Am J Clin Nutr. 39:100-104

Calabrese PP, Durett RT and Aquadro CF (2001) Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. Genetics 159:839-859

Caldwell E (2005) Molecular Evidence for Dietary Adaptation in Humans. The Centre for Genetic Anthropology, Department of Biology, University of London, London.

Caldwell E F (2004) Diet and the frequency of the alanine:glyoxylate aminotransferase Pro11Leu polymorphism in different human populations. Hum Genet 115:504-509

Castiglioni, A and G. Negro 1986 Fiuma di Pietra. Archivo della preistoria sahariana. Varese:Lativa

Cavalli-Sforza LL, Menozzi P and Piazza A (1994) The history and geography of Human Genes. Princeton NJ:Princeton University Press

Chaix,L and A. Grant 1987 – A study of a prehistoric population of sheep (ovis aries L) from Kerma (Sudan) Archaeozoologica 1 p93-107

Chapman, M (1993) Social and Biological aspects of Ethnicity. Oxford Science Publications. Oxford University Press, Oxford.

Chikhi L, Nichols RA, Barbujani G and Beaumont MA (2002) Y genetic data support the Neolithic demic diffusion model. PNAS 99(17):11008-11013

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V and Barbujani G (1998) Clines of nuclear DNA markers suggest a largely Neolithic ancestry of the European gene pool. Proc Natl Acad Sci 95:9053-9058

Clark A G Inference of Haplotypes from PCR-amplified Samples of Diploid Populations (1990)

Coehlo M, Luiselli D, Bertorelle G, Lopes AI, Seixas S, Destro-Bisol G and Rocha J (2005) Hum Genet 117:329-339

Clouarec D, Gouilloud S, Bornet F, Bruley des Varannes S, Bizais Y and Galmiche J-P (1991) Deficit en lactase et symptomes d'intolerance au lactose dans une population adulte saine originaire de l'ouest de la France. Gasteroenterol Clin Biol. 15:588-593

Cook G C (1979) Intestinal lactase status of adults in Papua New Guinea. Ann Hum Biol. 6(1):55-58

Close, A.E., and Wendorf, F. 1992. The beginnings of food production in the eastern Sahara. In Transitions to agriculture in prehistory. A.B. Gebauer and T.D. Price (eds) 63-72. Madison: Prehistory Press.

Clutton-Brock, J Chapter 3 'Cattle, Sheep and goats south of the Sahara: an archaeozoological perspective'. P30-37 – in Blench book

Cook G, Asp N, and Dahlqvist A (1973) Lactose absorption kinetics in Zambian African subjects. Br J Nutr 30:519-527.

Cook G and Howells G R (1968) Lactosuria in the African with Lactase deficiency. Am J Dig Dis. 13(7):634-637

Cook G and Kajubi S (1966) Tribal incidence of lactase deficiency in Uganda. Lancet 1:725-729.

Cook G, Lakin A, and Whitehead R (1967) Absorption of lactose and its digestion products in the normal and malnourished Ugandan. Gut 8:622-627.

Cook G and Al-Torki MT. (1975) High Intestinal lactase concentrations in adult Arabs in Saudi Arabia. Br Med J 3:135-6

Cox J and Elliott F (1974) Primary adult lactose intolerance in the Kivu lake area: Rwanda and the Bushi. Am J Dig Dis 19:714-724.

Craig O, Mulville J, Parker Pearson M, Sokol R, Gelsthorpe K, Stacey R and Collins M (2000) Detecting milk proteins in ancient pots. Nature 408:312

Craig O and Collins M J (2000) An improved method for the immunological detection of mineral bound protein using hydrofluoric acid and direct capture. J Imm Meth. 236:89-97

Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, and Underhill PA (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am.J.Hum Genet 70:1197-1214.

Cuatrecasas PF, Lockwood H and Caldwell J (1965) Lactase deficiency in the adult: a common occurrence. Lancet 1:14-18

Cuddenec Y, Delbruck H and Flatz G (1982) Distribution of the Adult Lactase Phenotypes-Lactose Absorber and Malabsorber – in a group of 131 Army Recruits. Gastroenterol Clin Biol. 6:776-779

Currat M and Excoffier L (2005) The effect of Neolithic expansion on European molecular diversity. Proc Biol Sci. 272)1564):679-88

Cummings MR (2000) Human Heredity:Principles and Issues 5th edition. Brooks/Cole, Pacific Grove.

Czeizel A, Flatz G and Flatz S D (1983) Prevalence of primary adult lactose malabsorption in Hungary. Am J Hum Genet 64:398-401

Dahlqvist A and Borgstrom B (1961) Digestion and absorption of disaccharidases in man. Biochem J 81:411-418

Dahlqvist A and Lindberg T (1966) Development of the intestinal disaccharidase and alkaline phosphatase activities in the human fetus. Clin Sci 30:507-512

Dales, G F (1973) Archaeological and radiocarbon chronologies for Protohistoric South Asia. Chpt 11 p: 157-170. South Asian Archaeology. Ed – Hammond, N. Gerald Duckworth and company ltd. Duckworth.

Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R (1990) Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics 6:575-7

De Ritis F, Balestrieri G G, Ruggiero G, Filosa E and Auricchio S (1970) High Frequency of Lactase Activity Deficiency in Small Bowel of Adults in the Neapolitan Area. Ensym. Biol. Clin. 11:263-267

Demoule J-P and Perles C (1993) The Greek Neolithic: a new review. Journal of World Prehistory 7:355-416

Desai H G, Gupte U V, Pradhan A G, Thakkar K D and Antia F P (1969) Ind J Med Sci. 730-736

Devlin B and Risch N (1995) A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. Genomics 29:311-322

Dill J E, Levy M, Wells R F and Weser E (1972) Lactase deficiency in Mexican-American males. Am J Clin Nutr 25:869-870

Doell RG and Kretchmer N (1962) Studies of small intestine during development. I. Distribution and activity of a-galactosidases. Biochim Biophys Acta 62:353-362

Dumayne-Preaty L (2001) Human impact on vegetation. Handbook of Archaeological Sciences. Chichester, Wiley: 379-392

Edmonds CA, Lillie AS and Cavalli-Sforza LL (2003) Mutations arising in the wave front of an expanding population. PNAS

Elbein A D, Pan Y T, Pastuszak I and Carroll D (2003) New insights on trehalose: a multifuntional molecule. Glycobiology 13(4):17-27

Ellegren H (2000) Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet 24:400-402

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, and Jarvela I (2002) Identification of a variant associated with adult-type hypolactasia. Nat.Genet. 30:233-237.

Enattah NS, Forsblom C, Raspinera H, Tuomi T, Groop P-H, Jarvela I and the FinnDiane Study Group (2004) The genetic variant of lactase persistence C (-13910) T as a risk factor for type I and type II diabetes in the Finnish population. Eur J Clin Nutr 58:1319-1322

Enattah NS, Valimaki VV, Valimaki MJ, Loyttyniemi E, Sahi T and Jarvela I (2004) Molecularly defined lactose malabsorption, peak bone mass and bone turnover rate in young finnish men. Calcif Tissue Int 75(6):488-93

Enattah NS, Pekkarinen T, Valimaki KJ, Loyttyniemi E and Jarvela I (2005) Genetically defined adult-type hypolactasia and self-reported lactose intolerance as risk factors of osteoporosis in Finnish postmenopausal women. Eur J Clin Nitr 59(10):1105-11

Excoffier (2004) Special Issue: Analytical methods in phylogeography and genetic structure. Mol evol 13:727

Excoffier L and Slatkin M. (1995) Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population. Mol Biol Evol 12(5):921-927

Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491

Fanso, V.G. (1989) Cameroon History for Secondary Schools and Colleges. Macmillan. Cameroon

Ferguson A, MacDonald D M and Brydon G W (1984) Prevalence of lactase deficiency in British adults. Gut 25:163-167

Fielding J, Harrington M, and Fottrell P (1981) The incidence of primary hypolactasia amongst the Irish. Ir J Med Sci 150:276-277.

Filali A, Ben Hassine L, Dhouib H, Matri S, Ben Ammar A and Garoui E (1987) Etude de la malabsorption du lactose par le test respiratoire a l'hydrogene dans une population de 70 adultes tunisiens. Gastroenterol Clin Bio. 11:554-557

Fischer A (1982) Trade in Danubian shaft-hole axes and the introduction of Neolithic economy in Denmark. Journal of Danish Archaeology 1:7-12

Fisher, K (1973) Archaeological field surveys in Afghan Seistan 1960-70. chpt 10 p.131-156 South Asian Archaeology. Ed – Hammond, N. Gerald Duckworth and company ltd. Duckworth.

Fix AG (1996) Gene frequency clines in Europe: demic diffusion or natural selection. J Roy Anthr Inst 2:625-43

Flannery,K. 1969 Origins and ecological effects of early domestication in Iran and the Near East. In The Domestication and exploitation of plants and animals. P.J.Ucko and G.W.Dimbleby (eds) 73-100 London:Duckworth.

Flatz G and Rotthauwe HW (1977) The Human Lactase polymorphism: physiology and genetics of lactase absorption and malabsorption. Prog Med Genet 2: 205 – 249

Flatz G (1984) Gene-dosage effect on intestinal lactase activity demonstrated in vivo. Am J Hum Genet 36:306-10.

Flatz G (1987) Genetics of lactose digestion in humans. Adv.Hum.Genet. 16:1-77.

Flatz G and Rotthauwe HW (1971) Evidence against nutritional adaptation to tolerance to lactase. Humangenetik 13:118-25

Flatz G and Rotthauwe HW (1973) Lactose Nutrition and Natural Selection. Lancet 76-77

Flatz G and Rotthauwe HW (1977) The human lactase polymorphism: physiology and genetics of lactose absorption and malabsorption. Progr Med Genet 2:205-249

Flatz G, Henze H J, Palabiyikoglu E, Dagalp K and Turkkan T (1986) Distribution of the Adult Lactase Phenotypes in Turkey. Trop geogr Med. 38:255-258

Flatz G, Howell J N, Doench J and Flatz S D (1982) Distribution of Physiological Adult Lactase Phenotypes, Lactose Absorber and Malabsorber, in Germany. Hum Genet. 62:152-157

Flatz G, Schildge C and Sekou H (1986) Distribution of Adult Lactase Phenotypes in the Tuareg of Niger. Am J Hum Genet. 38:515-520

Gautier,A. 1987 Prehistoric men and cattle in North Africa: a dearth of data and a surfeit of models. In Prehistory of arid North Africa. A.Close (ed), 163-87. Dallas: SMU Press.

Gautier,A 1980. Contributions to the archaeolozoology of Egypt. In Prehistory of the eastern Sahara, F.Wendorf & R.Schild (eds), 317-44. New York: Academic Press.

Gautier,A 1984. Archaeozoology of the Bir Kiseiba region, eastern Sahara. In cattle-keepers of the eastern Sahara: the Neolithic of Bir Kiseiba, A.E. Close (ed) 49-72 Dallas: SMU Press.

Gudmand-Hoyer A and Jarnum S (1969) Lactose malabsorption in Greenland Eskimos Acta Med Scand. 186:235-7

Gudmand-Hoyer E, Fenger HJ, Kern-Hansen P and Madsen PR (1987) Sucrase deficiency in Greenland: Incidence and genetic aspects. Scand J Gastroenterol. 22:24-28

Guo SW and Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361-372.

Gabriel S B, Schaffner S F, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero S N, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander E S, Daly M J and Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225-2229

Gabriel,B 1973 Steinplatze: Feuerstellen neolithischer nomaden in der Sahara. Libyca 21 151-8

Gabriel,B 1987 Paleoecological evidence from Neolithic fireplaces in the Sahara. African Archaeological Review 5 93-104

Geddes DS (1983) Neolithic transhumance in the Mediterranean Pyrenees. World Archaeology 15:51-66

Gendrel D, Dupont C, Richard-Lenoble D, Gendrel C, Nardou M and Chaussain M (1989) Milk Lactose Malabsorption in Gabon Measured by the Breath Hydrogen Test. J Pediatr Gastroenterol Nutr. 8(4):545-547

Gibney SFA, Munroe V and Nurse GT (1981) Lactose absorption in a Western Massim population. Ann Hum Biol. 8(5):477-480

Goldstein D B, Tate, S K and Sisodiya S M (2003) Pharmacogenetics goes Genomic. Nature Reviews 4:937-947

Grand R J, Montgomery R K, Chitkara D K and Hirschhorn J N (2003) Changing genes; losing lactase. Gut 52:617-619

Grigson,C – 1989. Size and sex: evidence for the domestication of cattle in the Near East. In The Beginnings of agricultures. A Milles, D Williams, N Gardner (eds) 77 – 109 BAR International Series 496. Oxford: BAR

Groot PC, Bleeker MJ (1989) The human alpha-amylase multigene family consists of haplotypes with variable numbers of genes. Genomics 5(1):29-42

Gusfield D (2001) Inference of Haplotypes from Samples of Diploid Populations: Complexity and Algorithms. J Comp Biol 8(3):305-323

Gudmand-Hoyer E and Jarnum S (1969) Lactose malabsorption in Greenland Eskimos. Acta Med Scand. 186:235-237

Guyton AC and Hall JE (1996) Textbook of Medical Physiology. P831-836Ninth Edition WB Saunders Company Philadelphia London Toronto Montreal Sydney Tokyo

Hahn M W, Rockman M V, Soranzo N, Goldstein D B and Wray G A (2004) Population Genetic and Phylogenetic Evidence for Positive Selection on Regulatory Mutations at the Factor VII Locus in Humans. Genet 167:867-877

Haldane JBS (1949) Disease and Evolution. Ric. Scient. 19(Suppl.1):3-10

Hamblin M T, Thompson E E and Di Rienzo A (2002) Complex Signatures of Natural Selection at the Duffy Blood Group Locus. Am J Hum Genet 70:369-383

Harlan, JR (1989) Wild-grass seed harvesting in the Sahara and Sub-Sahara of Africa. Foraging and farming: the evolution of plant exploitation. D.R.G.C.H. Harris. London, Unwin Hyman:79-98

Harvey CB (1994) The biochemical and genetical analysis of lactase phlorizin hydrolase: with specific reference to the lactase persistence/non-persistence polymorphism in man, Genetics and Biometry, University of London, London.

Harvey CB, Fox MF, Jeggo PA, Mantei N, Povey S and Swallow DM (1993) Regional localization of the lactase-phlorizin hydrolase gene, LCT, to chromosome 2q21. Ann Hum Genet 57:179-85

Harvey CB, Hollox EJ, Poulter M, Wang Y, Rossi M et al (1998) Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism. Ann Hum Genet 62:215-23

Harvey CB, Pratt WS, Islam I, Whitehouse DB, Swallow DM (1995) DNA polymorphisms in the lactase gene. Linkage disequilibrium across the 70-kb region. Eur. J. Hum Genet. 3:27-41

Harvey CB, Wang Y, Darmoul D, Phillips A, Mantei N and Swallow DM (1996) Characterisation of a human homologue of a yeast cell division cycle gene, MCM6, located adjacent to the 5' end of the lactase gene of chromosome 2q 21. FEBS Lett 398:135-40

Harvey CB, Wang Y, Hughes LA, Swallow DM, Thurrell WP, Sams VR, Barton R, Lanzon-Miller S and Sarner M (1995) Studies on the expression of intestinal lactase in different individuals. Gut 36:28-33

Harvey PH, Martin RD and Clutton-Brock TH (1987) Life Histories in Comparative Perspective. Chpt 16:181-196 in "Primate Societies" Eds. Smuts BB, Cheney DL, Seyfarth RM, Wrangham RW and Struhsaker TT. The University of Chicago Press, Chicago and London.

Hayes B J, Vissher P M, McPartlan H C and Goddard M E (2003) Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. Genome Res. 13:635-643

Hijaki SS, Abulaban A, Ammarin Z et al (1983) Distribution of adult lactase phenotypes in Bedouins and in urban and agricultural populations of Jordan. Trop Geogr Med. 35:157-161

Hughes AL and Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class 1 loci reveals overdominant selection. Nature 335(6168):167-70

Hussein L, Flatz SD, Kuhnan W et al (1982) Distribution of human adult lactase phenotypes in Egypt. Hum Hered. 32:94-99

Ho M W, Povey S and Swallow D (1982) Lactase Polymorphism in Adult British Natives: Estimating Allele Frequencies by Enzyme Assays in Autopsy Samples. Am J Hum Genet 34:650-657

Hogenauer, C, Hammer, H F, Mellitzer, K, Renner, W, and Toplak, H. Evaluation of a New Genetic Test Compared to the Lactose Hydrogen Breath Test for the Diagnosis of Acquired Primary Lactase Deficiency. Gastroenterlogy 124 Suppl1, A-64. 2003. Ref Type: Abstract

Holden C and Mace R (1997) Phylogenetic analysis of the evolution of lactase digestion in adults. Hum.Biol. 69:605-628.

Hollox EJ and Swallow DM (2002) chapter 14:Lactase Deficiency:Biological and Medical Aspects of the Adult Human Lactase Polymorphism. From "The Genetic Basis of Common Diseases" Eds. King, RA, Potter JI and Motulsky AG 2nd Edition. OUP. Oxford

Hollox EJ (2000) Molecular and population genetics analyses of variation within and surrounding the human lactase gene. MRC Biochemical Genetics Unit, University of London, London.

Hollox EJ, Poulter, M, Wang Y, Krause A, and Swallow DM (1999) Common polymorphism in a highly variable region upstream of the human lactase gene affects DNA-protein interactions. Eur J Hum Genet. 7:791-800

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, and Swallow DM (2001) Lactase haplotype diversity in the Old World. Am.J.Hum.Genet. 68:160-172.

Holzel A, Schwarz V and Sutcliffe KW (1959) Defective Lactose Absorption causing malnutrition in Infancy. Lancet 1126:1128

Howell J N, Mellmann J, Ehlers P and Flatz G (1980) Intestinal Disaccharidase Activities and Activity Ratios in a Group of 60 German Subjects. Hepato-Gastroenterol 27:208-212

Howell JN, Schockenhoff T, and Flatz G (1981) Population screening for the human adult lactase phenotypes with a multiple breath version of the breath hydrogen test. Hum Genet 57:276-278.

Howland J (1921) Prolonged intolerance to carbohydrates. Trans Am Pediatr Soc 33:11-19

Hussein L, Flatz SD, Kahnau W and Flatz G (1982) Distribution of human adult lactase phenotypes in Egypt. Hum Hered 32:94-99

Iqbal T H, Wood G M, Lewis K O, Leek J P and Cooper B T (1993) Prevalence of primary lactase deficiency in adult residents of west Birmingham. BMJ 306:1303

Jackes M, and Lubell D (1992) The early Neolithic human remains from Gruta do Caldeirao. In J Zilhao (ed) Gruta do Caldeirao: O Neolitico Antigo pp 259-95 Lisbon: Instituto Portugues do Patrimonio Arquitectonico e Arqueologico.

Jackson R T and Latham M C (1978) Lactose and Milk Intolerance in Tanzania. East Afr Med J. 55(7):

Jacob R, Weiner J R, Stadge S and Naim H Y (2000) Additional N-Glycosylation and Its Impact on the Folding of Intestinal Lactase-phlorizin Hydrolase. J Biol Chem 275(14):10630-10637

Jarvela I, Enattah NS, Kokkonen J, Varilo T, Savilahti E et al (1998). Assignment of the locus for congenital lactase deficiency to 2q21, in the vicinity of but separate from the lactase-phlorizin hydrolase gene. Am J Hum Genet 63:1078-85

Jenkins T, Gibney SFA, Nurse GT and Penketh RJA (1981) Persistent high intestinal lactase activity in Papua New Guinea. Lactose absorption curves in two populations. Ann Hum Biol. 8(5):477-451

Jenkins T, Lehmann H and Nurse G T (1974) Public Health and Genetic Consitution of the San ("Bushmen"):Carbohydrate Metabolism and Acetylator Status of the !Kung of Tsumkwe in the North-Western Kalahari. BMJ 2:23-26

Jennbert K (1984) Den Produktiva gavan:tradition och innovation I Sydskandinavien for omkring 5 300 ar sedan. Lund: Acta Archaeologica Lundensia 4

Jersky J and Kinsley R H (1967) Lactase Deficiency in the South African Bantu. S A Med J. 41:1194-1196

Johnson J D, Simoons F J, Hurwitz R, Grange A, Mitchell C H, Sinatra F R, Sunshine P, Robertson W V, Bennett P H and Kretchmer N (1977) Lactose malabsorption among the Pima Indians of Arizona. Gastroent. 73:1299-1304

Jorde L B (2000) Linkage Disequilibrium and the Search for Complex Disease Genes. Genome Res 10:1435-1444

Jussila J (1969) Milk Intolerance and Lactose Malabsorption in Hospital Patients and Young Servicemen in Finland. Ann of Clin Res 1:199-207

Jusilla J (1969) Diagnosis of Lactose Malabsorption by the Lactose Tolerance Test with Peroral Ethanol Administration. Scand J Gastroent 4:361-368

Jussila J, Isokoski M and Launiala K (1970) Prevalence of Lactose Malabsorption in a Finnish Rural Population. Scand J Gastroent. 5:49-56

Kaderali L, Deshpande A, Nolan J P and White P S (2003) Primer-design for multiplexed genotyping. Nuc Acids Res 31(6):1796-1802

Kanaghinis T, Hazioannon J, Deliarygri N et al (1974) Primary lactase deficiency in Greek adults. Am J Digest Dis. 19:1021-1027

Kar P and Tandon RK (1985) Lactose Intolerance in Nagaland. Ind J Med Res. 82:254-256

Kayser M, de Knijff P, Dieltjes P, Krawczak M, Nagy M, Zerjal T, Pandya A, Tyler-Smith C, and Roewer L (1997) Applications of microsatellite-based Y chromosome haplotyping. Electrophoresis 18:1602-1607.

Kayser M, Brauer S and Stoneking M (2003) A Genome Scan to Detect Candidate Regions Influenced by Local Natural Selection in Human Populations. Mol Biol Evol. 20(6):893-900

Keane R, O'Grady J G, Sheil J, Stevens F M, Egan-Mitchell B, McNicholl B, McCarthy C F and Fottrell P F (1983) Intestinal lactase, sucrase and alkaline phosphate in relation to age, sex and site of intestinal biopsy in 477 Irish subjects. J Clin Pathol. 36:74-77

Keusch GT, Troncale FJ, Thavarama B, Prinyanont P, Anderson PR, et al (1969) Lactase deficiency in Thailand: effect of prolonged lactose feeding. Am J Clin Nutr. 22:638-41

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624-626

Klein RG (1989) The human career: human and biological and cultural origins. London, University of Chicago Press.

Kozlov, A.I., Balanovskaya,E.V., Nurbaev,S.D., Balanovskaya,O.I. – 'Gene geography of primary hypolactasia in populations of the old world.' Russian Journal of genetics vol.34 no 4. p.455-454 1998

Kozlov A I (1998) Hypolactasia in the indigenous populations of Northern Russia. Int J Circum He. 57:18-21

Kreitman M (2000) Methods to detect selection in populations with applications to the human. Ann Rev Gen Hum Genet 01:539-59

Kretchmer N, Ransome-Kuti O, Hurwitz R, Dungy C, and Alakija W (1971) Intestinal absorption of lactose in Nigerian ethnic groups. Lancet 2:392-395.

Kruglyak S, Durrett RT, Schug MD and Aquadro CF (1998) Equilibrium distribution of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA 95:10774-10778

Kruse TA, Bolund L, Grzeschik KH, Ropers HH, Sjostrom H, Noren O, Mantei N and Semenza G (1988) The human lactase-phlorizin hydrolase gene is located on chromosome 2 FEBS Letts 240:123-126

Kuokkanen M, Enattah NS, Oksanen A, Savilahti E, Orpana A, and Jarvela I (2003) Transcriptional regulation of the lactase-phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. Gut 52:647-652.

Kurt I, Abou Ghoush M, Hasimi A, Serdar M, and Kutluay T (2003) Comparison of indirect methods of lactose absorption. Turk J Med Sci 33:103-110.

Kvetchmer, M. (1989) 'Invited Editorial: Expression of Lactase during development' Am J Hum Genet. 45: 487-8

Lember M, Tamm V and Villako K (1991) Lactose malabsorption in Estonians and Russians. Eur J Gastroenterol. Hematol 3:479-481

Lacey,S.W., Naim,H.Y., Magness,R.R., Gething,M.J., and Sambrook,J.F. (1994). Expression of lactase-phlorizin hydrolase in sheep is regulated at the RNA level. Biochem. J. 302, 929-935.

Lai Y and Sun F (2003) The Relationship Between Microsatellite Slippage Mutation Rate and the Number of Repeat Units. Mol Biol Evol 20(12):2123-2131

Larsen, CS (2000) Dietary reconstruction and nutritional assessment of past peoples: the bioanthropological record. The Cambridge World History of Food. K.F.K.a.K.C Ornelas. Cambridge, Cambridge University Press. 1:13-34

Lee,S.Y., Wang,Z., Lin,C.K., Contag,C.H., Olds,L.C., Cooper,A.D., and Sibley,E. (2002). Regulation of intestine-specific spatiotemporal expression by the rat lactase promoter. J. Biol. Chem. 277, 13099-13105.

Leichter J (1971) Lactose Tolerance in a Jewish Population. Dig Dis. 16(12):1123-1126

Leichter J (1972) Lactose Tolerance in a Slavic Population. Dig Dis. 17(1):73-76

Leichter J and Lee M (1971) Lactose Intolerance in Canadian West Coast Indians. Dig Dis. 16(9):809-813

Leis R, Tojo R, Pavon P and Douwes A (1997) Prevalence of Lactose Malabsorption in Galicia. J Pediatr Gastroenterol Nutr. 25:296-300

Legge, T – 1996. The beginning of caprine domestication in southwest Asia, In The Origins and spread of agriculture and pastoralism in Eurasia, D.R. Harris (Ed) 238-62. London UCL Press

Lember M, Tamm A, Piirsoo A et al (1995) Lactose malabsorption in Khants in western Siberia. Scand J Gastroenterol. 30:225-227

Lember M, Tamm A and Villako K (1991) Lactose Malabsorption in Estonians and Russians. Eur J Gasteroenterol and Hep. 3:479-481

Lewontin RC (1972) The apportionment of human diversity. Evol. Biol. 6:381-398

Lin S, Cutler D J, Zwick M E and Chakravarti A (2002) Haplotype Inference in Random Population Samples. Am J hum Genet 71:1129-1137

Lisker R, Gonzalez B, Daltabuit M (1975) Recessive inheritance of the adult type of intestinal lactase deficiency. Am J Hum Genet 27:662-64

Lisker R, Lopez-Habib G, Daltabuit M, Rostenberg I and Arroyo P (1974) Lactase deficiency in a rural area of Mexico. Am J Clin Nutr 27:756-759

Lloyd M, Mevissen G, Fischer M, Olsen W, Goodspeed D, Genini M, Boll W, Semenza G and Mantei N (1992) Regulation of Intestinal Lactase in Adult Hypolactasia. J Clin Invet 89:524-529

Lonjou C, Zhang W, Collins A, Tapper W J, Eiram E, Nikolas M and Morton N E (2003) Linkage Disequilibrium in human populations. PNAS 100(10) 6069-6074

Luikart G, England P R, Tallmon D, Jordan S and Taberlet P (2003) The Power and Promise of Population Genetics: from genotyping to genome typing. Nat Rev 4:981-994

Luyken R, Luyken-Koning and Immikhuizen M J T (1971) Lactose Intolerance in Surinam. Trop geogr med. 23:54-59

MacDonald, KC 'The Origins of African Livestock: indigenous or imported?" pges 2-7 Blench Book

Mackey AD, Henderson GN and Gregory JF (2002) Enzymatic Hydrolysis of Pyridoxine-5'-B-D-glucoside is Catalyzed by Intestinal Lactase-Phlorizin Hydrolase. J Biol Chem. 277(30):26858-26864

Madzarovova-Nohejlova J (1982) Small Bowel disaccharidase activity in Czech population and in gipsy population living in West Bohemia. 7th congress; organisation mondiale de gastroenterologie; Stockholm Abs.100

Mainguet P, Faille I, Destrebecq L, Devogelaer J-P and Nagant de Deuxchaisnes C (1991) Lactose intolerance, calcium intake and osteopenia. Lancet 338:

Maiuri L, Raia V, Potter J, Swallow D, Ho MW et al (1991) Mosaic pattern of lactase expression by villous enterocytes in human adult-type hypolactasia. Gastroenterology 100:359-69

Maiuri L, Rossi M, Raia V, Garipoli V, Hughes LA et al (1994) Mosaic regulation of lactase in human adult-type hypolactasia. Gastroenterology 107:54-60

Mallory JP (1989) In search of the Indo-Europeans. London: Thames and Hudson

Mantei N, Villa M, Enzler T, Wacker H, Boll W et al (1988) Complete primary structure of human and rabbit lactase-phlorizin hydrolase: implications for biosynthesis, membrane anchoring and evolution of the enzyme. EMBO J. 7:2705-13

Mathias N, Bayes M, and Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. Hum Mol.Genet 3:115-123.

McCrackern RD (1971) Lactase deficiency: An example of dietary evolution. Curr Anthropol. 12:479-500

McDonald JH and Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-654

McDonald J (1996) Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol Biol Evol. 13:235-260

McDonald J (1998) Improved tests for heterogeneity across a region of DNA sequence in ther ratio of polymorphism to divergence. Mol Biol Evol 13:377-384

McEvedy C (1995) The Penguin Atlas of African History. Penguin Reference Books, London.

McNair A, Gudmand-Hoyer E, Jarnum S and Orrild L (1972) Sucrose malabsorption in Greenland Brit Med J 2:19-21

Medjugorac I, Kustermann W, Lazar P, Russ I and Pirchner F (1994) Marker-derived phylogeny of European cattle supports demic expansion of agriculture. Anim Genet. Suppl 1:19-27

Menard D (1994) Development of human intestinal and gastric enzymes. Acta Paediatr Suppl 405:1-6

Metneki J, Czeizel A, Flatz S, Flatz G (1984) A study of lactose absorption capacity in twins. Hum Genet 67:296-300

Montgomery R K, Buller H A, Rings E H H M and Grand R J (1991) Lactose intolerance and the genetic regulation of intestinal lactase-phlorizin hydrolase. FASEB 5: 2824-2832

Morton N E, Zhang W, Taillon-Miller P, Ennis S, Kwok P-Y, and Collins A (2001) The optimal measure of allelic association. PNAS 98(9):5217-5221

Mulcare C A, Weale M E, Jones A L, Connell B, Zeitlyn D, Tarekegn A,. Swallow D M, Bradman N and Thomas M G (2004) The T allele of a SNP 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase persistence phenotype in Africans. Am J Hum Gen 74:1102-1110

Murdock G. 1967. Ethnographic Atlas. Pittsburgh PA: University of Pittsburgh Press.

Murray IA, Coupland K, Smith JA, Ansell D and Long RD (2000) Intestinal trehalase activity in a UK population: establishing a normal range and the effect of disease. Br J Nutr 83:241-245

Nachman MW and Crowell SL (2000) Estimate of the mutation rate per nucelotide in humans. Genetics 156:297-304

Naim HY (1992) Processing of human pro-lactase-phlorizin hydrolase at reduced temperatures: cleavage is preceded by complex glycosylation. Biochem J. 1:13-6

Naim HY, Jacob R, Naim H, Sambrook JF, Gething MJ (1994) The pro region of human intestinal lactase-phlorizin hydrolase. J Biol Chem 269:26933-43

Naim HY (1995) The pro-region of human intestinal lactase-phlorizin hydrolase is enzymatically inactive towards lactose. Biol Chem Hoppe-Seyler 376:255-58

Nasrallah S M (1979) Lactose intolerance in the Lebanese population and in "Mediterranean lymphoma". Am J Clin Nutr. 32:1994-1996

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nejati-Javaremi A and Smith C (1996) Assigning Linkage Haplotypes from Parent and Progeny Genotypes. Genet 142:1363-1367

Nemeth K, Plumb GW, Berrin J-G, Juge N, Jacob R, Naim H Y, Williamson G, Swallow DM and Kroon PA (2003) Deglycosylation by small intestinal epithelial cell B-glucosidases is a critical step in the absorption and metabolism of dietary flavonoid glycosides in humans. Eur J Nutr 42:29-42

Newcomer A and McGill DB (1966) Distribution of disaccharidase activity in the small bowel of normal and lactase deficient subjects. Gastroent. 51:481-488

Newcomer A, McGill DB, Thomas P, and Hofmann A (1975) Prospective comparison of indirect methods for detecting lactase deficiency. New Eng J Med 24:1232-1235.

Newcomer A D, Thomas P J, McGill D B and Hofmann A F (1977) Lactase deficiency: a common genetic trait of the American Indian. Gastroent. 72:234-237

Newcomer A D, Gordon H, Thomas P J and McGill D B (1977) Family studies of Lactase Deficiency in the American Indian. Gastroent. 73:985-988

Newman J. 1995. The peopling of Africa: a geographic interpretation. Newhaven and London: Yale University Press.

Nichols B L, Avery S, Sen P, Swallow D M, Hahn D and Sterchi E (2003) The maltase-glucoamylase gene: Common ancestry to sucrase-isomaltase with complementary starch digestion activities. PNAS 100:1432-1437

Niu T, Qin Z S, Xu X, Liu JS (2002) Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms. Am J Hum Genet 70:157-169

Nurse G and Jenkins T (1974) Lactose intolerance in San populations. Br Med J 2:728.

Oberholzer T, Mantei N and Semenza G (1993) The pro sequence of lactase-phlorizin hydrolase is required for the enzyme to reach the plasma membrane: an intramolecular chaperone? FEBS Lett 333:127-131

O'Keefe S and Adam J (1983) Primary lactose intolerance in Zulu adults. S Afr Med J 63:778-780.

O'Keefe S, Adam J, Cakata E, and Epstein S (1984) Nutritional support of malnourished lactose intolerant African patients. Gut 25:942-947.

Ohta T (1992) The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23:263-286

Olatunbosun D and Adadevoh B (1971) Lactase deficiency in Nigerians. Am J Dig Dis 16:909-914.

Olds LC and Sibley E (2003) Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Hum Mol.Genet 12:2333-2340.

Osier M V, Pakstis A J, Soodyall H, Comas D, Goldman D, Odunsi A, Okonofua F, Parnas J, Schulz L O, Bertranpetit J, Bonne-Tamir B, Lu R-B, Kidd J R and Kidd K K (2002) A Global Perspective on Genetic Variation at the ADH Genes Reveals Unusual Patterns of Linkage Disequilibrium and Diversity. Am J Hum Genet 71:84-99

Ott, J (1991) Analysis of human genetic linkage. The John Hopkins University Press, Baltimore and London.

Ouvendijk J, Peters WJ, Hollenberg CP, Ginsel JA, Fransen JA and Naim HY (1996) Congenital sucrase-isomaltase deficiency. Identification of a glutamine to proline substitution that leads to a transport block of sucrase-isomaltase in a pre-Golgi compartment. J Clin Invest 93:633-641

Pagnier J, Mears J G, Dunda-Belkhodja O, Schaefer-Rego K E, Beldjord C, Nagel R L and Labie D (1984) Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. Proc Natl Acad Sci 81:1771-1773

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, and Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am.J.Hum Genet 63:1839-1851.

Pena A S, Truelove S C and Whitehead R (1973) Morphology and Disaccharidase Levels of Jejunal Biopsy Specimens from Healthy British Subjects. Dig. 8:316-323

Petes TD (2001) Meiotic recombination hot spots and cold spots. Nature Rev. Genet. 2:360-369

Peukhuri K (2000) Lactose, Lactase, and Bowel Disorders. Institute of Biomedicine University of Helsinki, Helsinki.

Phillips M S, Lawrence R, Sachidanandam R, Morris A P, Balding D J, Donaldson M A, Studebaker J F, Ankener W M, Alfisi S V, Kuo F-S, Camisa A L, Pazorov V, Scott K E, Carey B J, Faith J, Katari G, Bhatti H A, Cyr J M, Derohannessian V, Elosua C, Forman A M, Grecco N M, Hock C R, Kuebler J M, Lathrop J A, Mockler M A, Nachtman E P, Restine S L, Varde S A, Hozza M J, Gelfand C A, Broxholme J, Abecasis G R, Boyce-Jacino M T, Cardon L R (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33:382-387

Pie S, Lalle J P, Blazy F, Laffitte J, Seve B and Oswald I P (2004) Weaning is Associated with an Upregulation of Expression of Inflammatory Cytokines in the Intestine of Piglets. Nutr Imm 641:647

Pietschmann P, Knoflach P and Wolosczczuk W (1991) Increased Serum Osteocalcin Levels in Patients with Lactase Deficiency. Am J Gastroent. 86(1):72-74

Plimmer RHA (1906) On the presence of lactase in the intestines of animals and on the adaptation of the intestine to lactose. J Physiol 35:20-31

Pluciennik M (1996) Genetics, archaeology and the wider world. Antiquity 70:13-14

Potter J, Ho M-W, Bolton H, Furth AJ, Swallow DM and Griffiths B (1985) Human Lactase and the Molecular Basis of Lactase Persistence. Biochem Genet 23(5/6):423-440

Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarner M, Korpela R, and Swallow DM (2003) The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. Ann.Hum.Genet. 67:298-311.

Price TG (ed) (2000) Europe's First Farmers. University of Wisconsin, Madison. Cambridge University Press.

Przeworski M (2002) The Signature of Positive Selection and Randomly Chosen Loci. Genetics 160:1179-1189

Puehkuri K. 2000. Lactose, lactase and bowel disorders, PhD thesis. Helsinki: Hakapaino.

Rab SM and Baseer A (1976) High Intestinal concentration in adult Pakistanis. Br Med J. 1:436

Rahimi AG, Delbruck H, Haeckel HW, Goedde H W and Flatz G (1976) Persistence if hugh intestinal lactase activity (lactose tolerance) in Afghanistan. Hum Genet 34:57-62

Ransome-Kuti O, Kretchmer N, Johnson J, and Gribble J (1972) Family studies of lactose intolerance in Nigerian ethnic groups. Pediatr Res 6:359.

Ransome-Kuti O, Kretchmer N, Johnson J, and Gribble J (1975) A genetic study of lactose digestion in Nigerian families. Gastroenterology 68:431-436.

Raspinera H, Savilahti E, Enattah N S, Kuokkanen M, Totterman N, Lindahl H, Jarvela I and Kolho K-L (2004) A genetic test which can be used to diagnose adult-type hypolactasia in children. Gut 53:1571-1576

Raspinera H, Forsblom C, Enattah N S, Halonen P, Salo K, Victorzon M, Mecklin J-P, Jarvinen H, Enholm S, Sellick G, Alazzouzi H, Houlston R, Robinson J, Groop P-H, Tomlinson I, Schwartz S Jr, Aaltonen L A, Jarvela I and The FinnDiane Study Group (2005) The C/C-13910 genotype of adult-type hypolactasia is associated with an increased risk of colorectal cancer in the Finnish population. Gut 54:643-647

Ray N, Currat M and Excoffier L (2003) Intra-Deme Molecular Diversity in Spatially Expanding Populations. Mol Biol Evol 20(1):76-86

Reed, C.A. (1977) Origins of agriculture. The Hague:Mouton

Renfrew AC (1987) Archaeology and Language: the puzzle of Indo-European Origins. London:Cape.

Renfrew C and Bahn P (1996) Archaeology: Theories, Methods and Practice. Thames and Hudson Ltd, London.

Reich D E and Goldstein D B (2001) Detecting Association in a Case-Control Study While Correcting for Population Stratification. Genet Epid 20:4-16

Richards M, Corte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H,Demaine A, Papiha S, Hedges R, Bandelt H-J and Sykes B (1996) Palaeolithic and Neolithic lineages in the European mitochondrial gene pool. Am J Hum Genet 59:185-203

Richards M (2003) The Neolithic Invasion of Europe. Ann Rev Anthropol. 32:135-62

Rinaldi E, Albini C, Costagliola et al (1984) High Frequency of lactose absorbers among adults with idiopathic senile and presnile cataract in a population with a high prevalence of lactose malabsorption. Lancet 1:255-357

Rings,E.H., de Boer,P.A., Moorman,A.F., van Beers,E.H., Dekker,J., Montgomery,R.K., Grand,R.J., and Büller,H.A. (1992). Lactase gene expression during early development of rat small intestine. Gastroenterology 103, 1154-1161.

Rings EH, Grand RJ and Buller HA (1994) Lactose Intolerance and lactase deficiency in children. Curr Opin Pediatr. 5:562-7

Rockman M V, Hahn M W, Soranzo N, Loisel D A, Goldstein D B and Wray G A (2004) Positive Selection on MMP3 Regulation Has Shaped Heart Disease Risk. Curr Biol 14:1531-1539

Rohman F and Nagano J (1903) *ber die Resorption und die ferntative Spaltung der Disaccharide im D*nndarm des augewachsenen Hundes. Arch Ges Physiol 95:533-605

Rooney AP and Zhang J (1999) Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive Darwinian selection? Mol. Biol. Evol. 16:552-569

Rosenberg n A, Pritchard J K, Weber J L, Cann H M, Kidd K K, Zhivotovsky L A and Feldman M W (2002) Genetic Structure of Human Populations. Science 298:2381-2385

Rosenkranz W, Hadorn B, Muller W, Heinz-Erian P, Hensen C and Flatz G (1982) Distribution of Human Adult Lactase Phenotypes in the Population of Austria. Hum Genet. 62:158-161

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, Armenteros M, Arroyo E, Barbujani G, Beckman G, Beckman L, Bertranpetit J,

Bosch E, Bradley DG, Brede G, Cooper G< Corte-Real HB, de Knijff P, Decorte R, Dubrova YE, Evgrafov O, Gilissen A, Glisic S, Golge M, Hill EW, Jeziorowska A, Kalaydjieva L, Kayser M, Kivisild T, Kravchenko SA, Krumina A, Kucinskas V, Lavinha J, Livshits LA, Malaspina P, Maria S, McElreavey K, Meitinger TA, Mikelsaar AV, Mitchell RJ, Nafa K, Nicholson J, Norby S, Pandya A, Parik J, Patsalis PC, Pereira L, Peterlin B, Pielberg G, Prata MJ, previdere C, Roewer L, Rootsi S, Rubinsztein DC, Saillard J, Santos FR, Stefanescu G, Sykes BC, Tolun A, Villems R, Tyler-Smith C and Jobling MA (2000) Y-Chromosomal diversity in Europe is clinal and influenced primarily by geography rather than by language. Am J Hum Genet 67(6):1526-43

Rotthauwe HW, El-Schallah M O and Flatz G (1971) Lactose intolerance in Arabs. Hum Genet 13:344-346.

Rotthauwe HW and Emons D (1971) Comparative study of Intestinal Disaccharidases in Thai and European subjects. Jg Heft 8:503-504

Rowley-Conwy P (1987) Animal bones in Mesolithic studies: recent progress and hopes for the future. In P.Rowley-Conwy, M Zvelebil and HP Blankholm (eds) Mesolithic Northwest Europe: recent trends pp 74-81. Sheffield: University of Sheffield.

Ruhlen M (1991) A Guide to the World's languages. London, England: Edward Arnold.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R and Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837

Sadre M and Karbasi K (1979) Lactose intolerance in Iran. Am J Clin Nutr 32:1948-1954

Sahi T, Isokoski M, Jussila J, Launiala K and Pyorala K (1973) Recessive inheritance of adult-type lactose malabsorption. Lancet 823-826

Sahi T (1974) The inheritance of selective adult-type lactose malabsorption. Scand J Gastroenterol 9 (Suppl-30):1-73.

Sahi T (1974) Lactose Malabsorption in Finnish-speaking and Swedish-speaking Populations in Finland. Scand J Gastroent. 9:303-308

Sahi T and Launiala K (1978) Manifestation and Occurrence of Selective Adult-Type Lactose Malabsorption in Finnish Teenagers. Dig Dis 23(8):699-704

Sahi T, Launiala K and Laitinen H (1983) Hypolactasia in a Fixed Cohort of Young Finnish Adults. A follow up Study. Scand J Gastroenterol. 18:865-870

Sahi T, Jussila J, Penttila IM, Sarna S and Isokoski M (1977) Serum lipids and proteins in lactose malabsorption. Am J Clin Nutr 30:476-481

Sanae H A, Saldanha W, Sugathan T N and Molla A M (2003) Comparison of Lactose Intolerance in Healthy Kuwaiti and Asian Volunteers. Med Princ Prac. 12:160-163

Saunders N J, Moxon E R and Gravenor M B (2003) Mutation rates:estimating phase variation rates when fitness differences are present and their impact on population structure. Microbiol. 149:485-495

Scarre, C and Fagan, BM (1997) Ancient Civilisations. Longman publishing group.

Schlotterer C (2002) A Microsatellite-Based Multilocus Screen for the Identification of Local Selective Sweeps. Genetics 160:753-763

Schneider S, Roessli D, Excoffier L (2002) Arlequin ver.2.000: A software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Sebastio G, Villa M, Sartorio R, Guzetta V, Poggi V et al (1989) Control of lactase in human adult-type hypolactasia and in weaning rabbits and rats. Am J Hum Genet 45:489-97

Segal JJ (1980) Hypothesis. Is Lactose a Dietary Risk Factor for Ischaemic Heart Disease? Int J Epid 9(3):271-6

Segal I, Gagjee P, Essop A, and Noormohamed A (1983) Lactase deficiency in the South African black population. Am J Clin Nutr 38:901-905.

Semino O, Passarinoa G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA (2000) The Genetic Legacy of Paleolithic Homo Sapiens Sapiens in Extant Europeans: A Y Chromosome Perspective. Sci. 290:1155-1159

Senewiratne B, Thambipillai S and Perera H (1977) Intestinal Lactase Deficiency in Ceylon (Sri Lanka). Gastroent 72(6):1257-1259

Shete S (2003) A note on the Optimal Measure of Allelic Association. Ann Hum Genet 67:189-191

Shostak M (1994) Lisa: The Life and Words of a !Kung Woman. 2nd edition. Earthscan Publications Ltd, London.

Simoons FJ (1970) Primary adult lactose intolerance and the milking habit: A problem in biological and cultural interrelations II. A culture historical hypothesis. Am J Dig Dis 15:695-710

Simoons FJ (1978) the geographic hypothesis and lactose malabsorption: a weighing of the evidence. Am J Dig Dis 23:963-980

Simoons FJ (1982) A Geographic Approach to Senile Cataracts: Possible links with Milk Consumption, Lactase Activity, and Galactose Metabolism. Dig Dis and Sci 27(3):257-263

Simoni L, Calafell F, Pettener D, Bertranpetit J and Barbujani G (2000) Geographic Patterns of mtDNA Diversity in Europe. Am J Hum Genet 66:262-278

Skovbjerg,H., Norén,O., Sjöström,H., and Danielsen,E.M. (1982). Further characterization of intestinal lactase-phlorizin hydrolase. Biochim. Biophys. Acta 707, 89-97.

Slatkin M (1995) A measure of population sbdivision based on microsatellite allele frequencies. Genetics 139:457-462

Slatkin M (2001) Simulation genealogies of selected alleles in a population of variable size. Genet Res Cambridge 78:49-57

Slatkin M and Bertorelle G (2001) The use of intraallelic variability for testing neutrality and estimating population growth rate. Genetics 158-865-74.

Snook C R, Mahmoud J N and Chang W P (1976) Lactose tolerance in adult Jordanian Arabs. Trop geogr Med. 28:333-335

Socha J and Ksiazyk J (1984) Prevalence of primary adult lactose malabsorption in Poland. Ann Hum Biol. 11(4):311-316

Sokal RR, Oden NL and Wilson C (1991) Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nat 351:97-98

Sowers M F and Winterfeldt E (1975) Lactose intolerance among Mexican Americans. Am J Clin Nutr. 28:704-705

Spanidou E P and Petrakis N L (1972) Lactose Intolerance in Greeks. Lancet 872-873

Stephens JC, Reich DE (1998) Dating the origin of the CCR5-Delta 32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62(6):1507-15

Stephens J C, Schneider J A, Tanguay J C, Acharya T et al (2001) Haplotype Variation and Linkage Disequilibrium in 313 Human Genes. Science 293:489-493

Stephens M and Donnelly P (2003) A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. Am J Hum Genet. 73:1162-1169

Stumpf M P H and McVean G A T (2003) Estimating Recombination Rates from Population Genetic Data. Nat Rev 4:959-968

Stumpf M P H and Goldstein D B (2003) Demography, Recombination Hotspot Intensity and the Block Structure of Linkage Disequilibrium. Curr Biol 13:1-8

Sung J-L and Shih P L (1972) The jejunal disaccharidases activity and lactose intolerance of Chinese adults. Asian J of Med 8:149-151

Swallow DM and Harvey CB (1993) Genetics of adult-type hypolactasia. Dyn.Nutr.Res. 3:1-7.

Swallow DM, Hollox EJ (2000) Chapter 76 The genetic polymorphism of intestinal lactase activity in adult humans. In Scriver CR et al (eds) The Metabolic and Molecular Basis of Inherited Disease, 8 ed. New York: McGraw-Hill.

Swallow,D.M. and Hollox,E.J. (2000). The genetic polymorphism of intestinal lactase activity in adult humans. In The Metabolic and Molecular Basis of Inherited Disease, C.R.Scriver, A.L.Beaudet, S.S.Sly, and D.Valle, eds. McGraw-Hill), pp. 1651-1562.

Swallow DM, Poulter M and Hollox EJ (2001) Intolerance to Lactose and Other Dietary Sugars. DMD 29(4) part2:513-516

Swallow DM (2003) Genetics of lactase persistence and lactose intolerance. Ann Rev Genet 73:197-219

Tadesse,K., Leung,D.T., and Yuen,R.C. (1992). The status of lactose absorption in Hong Kong Chinese children. Acta Paediatr. 81, 598-600.

Taillon-Miller P, Bauer-Sardina I, Saccone N L, Putzel J, Laitinen T, Cao A, Kere J, Pilia G, Rice J P and Kwok P-Y (2002) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat Genet 25:324-328

Tandon RK, Joshi YK, Singh DS et al (1981) Lactose intolerance in North and South Indians. Am J Clin Nutr 34:943-946

Tharpar, BK (1973) New traits of the Indus civilisation at Kalibangari: an appraisal. Chpt 7 p.85-104. South Asian Archaeology. Ed – Hammond, N. Gerald Duckworth and company ltd. Duckworth.

The Y chromosome consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. 12:339-348.

Thomas MG, Bradman N, and Flinn HM (1999) High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. Hum Genet 105:577-581.

Tishkoff S A, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma E H, Williams S M and Clark A G (2001) Haplotype Diversity and Linkage Disequilibrium at Human G6PD:Recent Origin of Alleles That Confer Malarial Resistance. Science 293:455-462

Toomajian C and Kreitman M (2002) Sequence Variation and Haplotype Structure at the Human HFE Locus. Genetics 161:1609-1923

Toomajian C, Ajioka R S, Jorde L B, Kushner J P and Kreitman M (2003) A Method for Detecting Recent Selection in the Human Genome From Allele Age Estimates. Genetics 165:287-297

Torp,N., Rossi,M., Troelsen,J.T., Olsen,J., and Danielsen,E.M. (1993). Lactase-phlorizin hydrolase and aminopeptidase N are differentially regulated in the small intestine of the pig. Biochem. J. 295, 177-182.

Troelsen JT (2005) Regulation of lactase expression. Implications for adult-type hypolactasia. Biochim Biophysi Acta

Troelsen JT, Olsen J, Moller J, and Sjostrom H (2003) An upstream polymorphism associated with lactase persistence has increased enhacer activity. Gastroenterology 125/6:1686-1694.

Troelsen JT, Mitchelmore C and Olsen J (2003) An enhancer activates the pig lactase phlorizin hydrolase promoter in intestinal cells. Gene 305(1):101-11
Troy CS, MacHugh DE, Bailey JF, Magee DA, Loftus RT, Cunningham P, Chamberlain AT, Sykes BC and Bradley DG (2001) Genetic evidence for Near-Eastern origins of European Cattle. Nat

Ulijaszek SJ, and Strickland SS (1993) Nutritional anthropology: prospects and perspectives. London, Smith-Gordon.

Varilo T, Paunio T, Parker A, Perola M, Meyer J, Terwiliger JD and Peltonen L (2003) The interval of linkage disequilibrium (LD) detected with microsatellite and

SNP markers in chromosomes of Finnish populations with different histories. Hum Molec genet 12(1):51-59

Veitch AM, Kelly P, Segal I, Soies SK and Farthing MJ (1998) Does sucrase deficiency in black South Africans protect against colonic disease? Lancet 351:1503-7

Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291:1304-1351

Vermeersch, P.M., P van Peer, J. Moeyersons, W Van Neer (1996) Neolithic occupation of the Sodmein area, Red Sea mountains, Egypt. Pwiti and Soper (1996) 411-19 – funny ref?

Verburg M, Renes IB, Van Nispen DJ, Ferdinandusse S, Jorritsma M, Buller HA, Einerhand AW and Dekker J (2002) Specific responses in rat small intestinal epithelial mRNA expression and protein levels during chemotherapeutic damage and regeneration. J Histochem Cytochem. 50(11):1525-36

Villa M., Brunschwiler,D., Gachter,T., Boll,W., Semenza,G., and Mantei,N. (1993). Region-specific expression of multiple lactase-phlorizin hydrolase genes in intestine of rabbit. FEBS Lett. 336, 70-74.

Villa S, Guiscafre H, Martinez H, Munoz O and Gutierrez G (1999) Seasonal diarrhoeal mortality among Mexican children. Bull World Health Organ 77(5):375-80

Wang Y, Harvey C, Rousset M and Swallow DM (1994) Expression of Human Intestinal mRNA Transcripts during Development: Analysis by a Semiquantitative RNA Polymerase Chain Reaction Method. Ped Res 36(4):514-520

Wang Z, Fang R, Olds L C and Sibley E (2004) Transcriptional regulation of the lactase-phlorizin hydrolase promoter by PDX-1. Am J Physiol Gastrointest Liver physiol 287:G555-G561

Wang Y, Harvey CB, Hollox EJ, Phillips AD, Poulter M, Clay P, Walker-Smith JA and Swallow DM (1998) The genetically programmed down-regulation of lactase in children. Gastroenterology 114:1230-36

Wang Y, Harvey CB, Pratt WS, Sams VR, Sarner M et al (1995) The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element Hum Mol Genet 4:657-62

Wang Y, Yan Y, Xue J et al (1984) Prevalence of primary lactose malabsorption in three populations of Northern China. Hum Genet. 67:103-106

Weale M E, Depondt C, MacDonald S J, Smith A, Lai P S, Shorvon S D, Wood N W and Goldstein D B (2003) Selection and Evaluation of Tagging SNPs in the Neuronal-Sodium-Channel Gene SCN1A: Implications for Linkage-Disequilibrium Gene Mapping. Am J Hum Genet 73:551-565

Weale ME, Weiss DA, Jager RF, Bradman N and Thomas MG (2002) Y Chromosome Evidence for Anglo-Saxon Mass Migration. Mol. Biol. Evol. 19(7):1008-1021

Weber JL and Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2(8):1123-1128

Welsh JD, Zschiesche OM, Willits VL and Russell L (1968) Studies of Lactose Intolerance in Families. Arch Intern Med 122:315-317

Welsh JD, Russell LC and Walker AW (1974) Changes in intestinal lactase and alkaline phosphate activity levels with age in the baboon (Papio papio). Gastroenterology 66:993-997

Wen C-P, Antonowicz I, Tovar E, McGandy RB and Gershoff SN (1973) Lactose feeding in lactose-intolerant monkeys. Am J Clin Nutr 26:1224-1228

Wendorf et al (1996) A late neolithic megalithic complex in the eastern Sahara: a preliminary report. In Interregional contacts in the later prehistory of northeastern Africa, L.Krzyzaniak, K.Kroeper, M.Kobusiewicz (eds) 152-32. Poznan: Poznan Archaeological Museum.

Wier BS and Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370

Wilson JF, Weiss DA, Richards M, Thomas MG, Bradman N and Goldstein DB (2000) Genetic evidence for different male and female roles during cultural transitions in the British Isles. PNAS 98(9): 5078-5083
Wiuf C (2001) Do delta F508 heterozygotes have a selective advantage? Genet Res 78(1):41-47

Woteki C E, Weser E and Young E A (1977) Lactose malabsorption in Mexican-American adults. Am J Clin Nutr. 30:470-475

Wright S (1931) Evolution in Mendelian populations. Genetics. 16:97-159
Wu R, Ma C-Z, Casella G (2001) Joint Linkage and Linkage Disequilibrium Mapping of Quantitative Trait Loci in Natural Populations. Genet 160:779-792

Xiao FX, Yotova V, Zietkiewicz E, Lovell A, Gehl D, Boureois S, Moreau C, Spanaki C, Plaitakis A, Moisan JP and Labuda D (2004) Human X-chromosomal

lineages in Europe reveal Middle Eastern and Asiatic contacts Eur J Hum Genet 12(4):301-11

Xu H and Fu Y-X (2004) Estimating Effective Population Size or Mutation Rate with Microsatellites. Genetics 166:555-563

Yap I, Berris B, Kang J Y, Math M, Chu M, Miller D and Pollard A (1989) Lactase deficiency in Singapore-Born and Canadian-Born Chinese. Dig Dis Sci. 34(7):1085-1088

Yang Z (2001) Adaptive molecular evolution. Handbook of statistical genetics, DJ Balding. London, John Wiley & Sons.

Yongfa A, Yongshan Y, Jiujin X, Ruofu D, Flatz SD, Kuhnau W and Flatz G (1984) Prevalence of primary adult lactose malabsorption in three populations of Northern China. Hum Genet. 67:103-106

Yoshida Y, Sasaki G, Goto S, Yanagiya S and Takashina K (1975) Studies on the etiology of milk intolerance in Japanese adults. Gastroent Jap. 10(1):29-34

Zhang J, Webb DM (2002) Accelerated protein evolution and origins of human-specific features:FOXP2 as an example. Genetics 162(4):1825-35

Zhivotovsky L A, Goldstein D B and Feldman M W (2001) Genetic Sampling Error of Distance $(\delta\mu)^2$ and Variation in Mutation Rate Among Microsatellite Loci. Mol Biol Evol 18(12):2141-2145

Zografos N, Kanaghinis T, Hatzioannou I and Gardikas C (1973) Lactose Intolerance in Greeks. Lancet 367

Appendices

A. STATISTICAL APPENDICES

A1: *GenoPheno script (written by Dr. Mike Weale)*
A2: *Graphing programme for Syssiphos output (written by Dr. Mark Thomas)*

B. DATA APPENDICES

B1: *Polymorphic loci defining the core lactase haplotypes (adapted from Hollox et al 2001)*
B2: *Global distribution of -13.9kb*T frequencies*
B3: *Global distribution of lactase persistence frequencies*

C. PAPER APPENDICES

C1: *Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarner M, Korpela R, and Swallow DM (2003) The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. Ann.Hum.Genet. 67:298-311.*
C2: *Mulcare C A, Weale M E, Jones A L, Connell B, Zeitlyn D, Tarekegn A,. Swallow D M, Bradman N and Thomas M G (2004) The T allele of a SNP 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase persistence phenotype in Africans. Am J Hum Gen 74:1102-1110*

Appendix A – Statistical procedures

A1). A detailed description of Dr. Mike Weale's procedure for the statistical comparison between phenotype and genotype data:

The procedure was as follows. (1) A value for p was drawn from a Beta(T+1, C+1) distribution, where T is the number of T alleles and C is the number of C alleles found in the genotyped group. This Beta distribution describes the posterior probability distribution for p having observed the genotype data, assuming a Uniform(0,1) prior. (2) The predicted frequency of true lactase persistence in the population, Ltrue, was calculated as $p^2+2p(1-p)$ (i.e. the expected frequency of TT+CT genotypes under Hardy-Weinberg equilibrium). (3) Values for fn and fp were drawn from Beta(11,107) and Beta(6,69) distributions respectively if phenotyping was by the blood glucose method, and from Beta(10,124) and Beta(6,116) distributions respectively if phenotyping was by the breath hydrogen method (again, these Beta distributions describe the posterior probabilities for fn and fp having observed the combined false error rate data reported above and assuming a Uniform(0,1) prior). (4) The predicted frequency of apparent lactose digesters accounting for phenotyping error, Lapp, was calculated as Ltrue (1-fp) + (1-Ltrue) FN. (5) A simulated value for nL, the number of lactose digesters observed in the phenotyped group was drawn from a Binomial (n, Lapp) distribution, where n is the number sampled in the phenotyped group. (6) Steps 1-5 were repeated N=100,000 times to build up a Monte Carlo sampling distribution for nL under the null hypothesis (that C/T genotype and phenotyping error alone account for the apparent frequency of lactose digesters). (7) Let S_g be the sum of simulated nL values greater than or equal to the observed nL value, and let S_l be the sum of simulated nL values less than or equal to the observed nL value. A two-tailed P-value for the observed nL under the null hypothesis was found as $2*min(S_g,S_l)/N$.

A2). 'R' code, written by Dr.Mark Thomas to create graphic output for Sisyphos program:

```
# 19-7-04 - WILL NOW WORK OUT NUMBER OF VALUSE OF S AND R IN LIKELIHOOD INPUT FILE

minusml <- 2   # PLOT WITHIN WHAT OF MAXIMUM LIKELIHOOD?:

rrsl <-read.table("OUTPUTFILE.like", header=TRUE);   # NAME OF FILE CONTAINING LIKELIHOOD VALUES
gvec <- rrsl[,1];
svec <- rrsl[,2];
len= length(gvec)
ng <- 1 + sum(0<diff(gvec))
ns <- len/ng
lmat <- matrix(rrsl[,3],ncol=ns,byrow=T);
startg <- rrsl[1,1]
starts <- rrsl[1,2]
endg <-  rrsl[len,1]
ends <-  rrsl[len,2]

for (i in 1: ng) {
for (j in 1: ns) {
if (lmat[i,j] < (-minusml+max(lmat))) lmat[i,j] <- (-minusml+(max(lmat)))
}
}
mxl<-max(lmat);
mnl<-min(lmat);

# TO MAKE A CONTOUR PLOT
x11()

image(seq(startg,endg,length=ng), seq(starts,ends,length=ns), lmat, zlim=c(mnl,mxl),
nlevels = 110,col = heat.colors(50),xlab="Growth",ylab="Selection",box=TRUE);

contour(seq(startg,endg,length=ng), seq(starts,ends,length=ns), add=TRUE,lmat,
zlim=c(mnl,mxl), axes=TRUE, shade = 0.95, col =
4,main="",xlab="Growth",ylab="Selection",box=TRUE, zlab="Likelihood");

# TO MAKE A PERSPECTIVE PLOT

x11()

persp(seq(startg,endg,length=ng), seq(starts,ends,length=ns), lmat, zlim=c(mnl,mxl),theta
= 45, phi = 30, expand = 0.9,nticks = 5,ticktype = "detailed", axes=TRUE, shade = 0.95,
col = 4,main="",xlab="Growth",ylab="Selection",box=TRUE, zlab="Likelihood");
```

Appendix B – Data

B1). Classification of core lactase haplotypes adapted from Hollox et al 2001.

	Polymorphism										
Core Lactase Haplotype	C-958T	A-946G	C-942G	TC-942/3 ∇∇	G-875A	A-678G	$A_8$552 / -559 A_9	C458T	G666A	T5579C	TG 6263/7 ∇∇
A	C	A	C	TC	G	A	A_9	C	G	C	TG
B	T	A	C	TC	G	A	A_8	C	A	T	∇∇
C	C	A	C	TC	G	G	A_8	C	G	T	TG
D	T	A	C	TC	A	A	A_8	C	A	T	∇∇
E	C	A	C	TC	G	G	A_8	C	G	C	TG
F	T	A	C	TC	G	A	A_8	C	A	C	∇∇
G	T	A	C	TC	G	A	A_8	C	A	T	TG
H	C	A	C	TC	G	A	A_9	C	G	T	TG
I	C	A	C	TC	G	A	A_9	C	G	T	∇∇
J	C	A	C	TC	G	A	A_8	C	G	C	TG
K	C	A	C	TC	G	A	A_8	C	G	T	TG
L	T	A	C	TC	G	A	A_8	C	G	T	TG
M	C	A	C	TC	G	G	A_8	C	G	T	TG
N	T	A	G	TC	G	A	A_8	C	A	T	∇∇
O	C	A	C	∇∇	G	A	A_9	T	A	T	TG
P	C	A	C	TC	G	A	A_8	C	A	T	∇∇
Q	C	A	C	TC	G	A	A_9	C	G	T	TG
R	T	A	C	∇∇	G	A	A_9	C	A	T	∇∇
S	C	A	C	∇∇	G	A	A_9	C	A	T	∇∇
T	C	A	C	∇∇	G	A	A_9	C	G	T	TG
U	C	A	C	∇∇	G	A	A_9	C	A	T	TG
V	C	G	C	∇∇	G	A	A_9	C	A	C	TG
W	C	A	C	TC	G	A	A_9	C	A	T	TG
X	C	A	C	∇∇	G	A	A_9	C	A	C	TG
Y	C	A	C	TC	G	A	A_8	C	A	T	TG
Z	C	A	C	TC	G	A	A_8	C	G	T	∇∇
a	T	A	C	TC	G	A	A_8	C	G	C	TG
b	C	A	C	∇∇	G	A	A_9	C	G	T	∇∇
c	C	A	C	TC	G	G	A_8	C	A	T	∇∇
d	C	A	C	TC	G	A	A_8	C	A	C	TG
e	C	A	G	TC	G	A	A_8	C	A	T	
f	C	A	C	∇∇	G	A	A_9	T	A	T	∇∇
g	C	A	C	∇∇	G	A	A_8	C	A	T	TG
h	C	A	C	TC	G	G	A_8	C	A	C	TG
i	T	A	C	TC	G	A	A_8	C	G	T	∇∇
j	C	A	C	TC	G	G	A_9	C	G	T	TG
k	C	A	C	TC	G	G	A_9	C	G	T	∇∇
l	C	A	C	TC	G	G	A_9	C	A	T	∇∇
m	C	A	C	∇∇	G	A	A_8	C	A	C	TG
n	C	A	C	∇∇	G	A	A_9	T	A	C	TG

B2 – GLOBAL DISTRIBUTION OF–13.9kb*T ALLELE FREQUENCIES

POPULATION	COUNTRY	LATITUDE	LONGITUDE	NO OF CHROMOSOMES	FREQUENCY -13.9kb t	SOURCE
Wolof	Senegal	14	-17	138	0.000	TCGA
Manjak	Senegal	14	-16	186	0.000	TCGA
Berber	Morocco	34	-3	154	0.136	TCGA
Fulani	Cameroon	6	11	98	0.112	TCGA
Hausa	Cameroon	6	11	36	0.139	TCGA
Nso	Cameroon	6	11	252	0.000	TCGA
Mambila	Cameroon	6	11	244	0.004	TCGA
Yamba	Cameroon	6	11	42	0.000	TCGA
Nuer	Sudanese	4	31	26	0.000	TCGA
Dinka	Sudanese	4	31	68	0.000	TCGA
Shaigi	Sudanese	15	32	22	0.000	TCGA
Ga'ali	Sudanese	15	32	60	0.000	TCGA
Bantu	Uganda	0	32	44	0.000	TCGA
Bantu	Malawi	-13	33	410	0.000	TCGA
Nuer	Ethiopian	8	34	238	0.000	TCGA
Anuak	Ethiopian	8	34	216	0.000	TCGA
Roma	Czech	50	14	162	0.099	Galton
Indian	North India	28	77	128	0.188	Galton
Indian	South India	13	80	68	0.132	Galton
San	Kalahari	-22	17	30	0.000	Galton
Bantu	South Africa	-26	28	50	0.000	Galton
Irish	Ireland	54	-7	65	0.954	TCGA
Algerian	Algeria	34	-1	21	0.333	TCGA
UK	London	51	0	64	0.734	TCGA
German	Germany	53	9	60	0.556	TCGA
Ashkenazi	East Europe	52	21	96	0.083	TCGA
Armenian	Armenia	40	44	88	0.011	TCGA
Kuwaiti	Kuwait	29	47	28	0.000	TCGA
Amharic	Ethiopian	9	38	119	0.000	TCGA
Orcadian	Orkney Islands	58	-3	32	0.688	Bersaglieri et al 2004
French Basque	France	43	-1	48	0.667	Bersaglieri et al 2004

French	France	48	2	58	0.431	Bersaglieri et al 2004
Sardinian	Italy	39	9	56	0.071	Bersaglieri et al 2004
North Italian	Italy	45	9	28	0.357	Bersaglieri et al 2004
Tuscan	Italy	43	11	16	0.063	Bersaglieri et al 2004
Scandinavians	Finland and Sweden	59	18	360	0.815	Bersaglieri et al 2004
Bedouin	Israel	31	34	98	0.031	Bersaglieri et al 2004
Druze	Israel	33	35	96	0.021	Bersaglieri et al 2004
Palestinian	Israel	31	35	102	0.039	Bersaglieri et al 2004
Russian	Russia	55	37	50	0.240	Bersaglieri et al 2004
Adygei	Russian Caucasus	44	42	34	0.118	Bersaglieri et al 2004
Hazara	Pakistan	35	66	50	0.080	Bersaglieri et al 2004
Sindhi	Pakistan	25	68	50	0.320	Bersaglieri et al 2004
Pathan	Pakistan	35	72	50	0.300	Bersaglieri et al 2004
Burusho	Pakistan	37	72	50	0.100	Bersaglieri et al 2004
Brahui	Pakistan	33	73	50	0.340	Bersaglieri et al 2004
Balochi	Pakistan	33	73	50	0.360	Bersaglieri et al 2004
Makrani	Pakistan	33	73	50	0.340	Bersaglieri et al 2004
Kalash	Pakistan	33	73	50	0.000	Bersaglieri et al 2004
Cambodian	Cambodia	11	104	22	0.000	Bersaglieri et al 2004
Miaozu	China	26	107	20	0.000	Bersaglieri et al 2004
Han	China	19	109	90	0.000	Bersaglieri et al 2004
Tujia	China	39	116	20	0.000	Bersaglieri et al 2004
Yizu	China	39	116	20	0.000	Bersaglieri et al 2004
Oroqen	China	39	116	20	0.000	Bersaglieri et al 2004
Daur	China	39	116	20	0.000	Bersaglieri et al 2004
Mongola	China	39	116	20	0.100	Bersaglieri et al 2004
Hezhen	China	39	116	20	0.000	Bersaglieri et al 2004
Xibo	China	39	116	18	0.000	Bersaglieri et al 2004
Uygur	China	39	116	20	0.050	Bersaglieri et al 2004

Dai	China	39	116	20	0.000	Bersaglieri et al 2004
Lahu	China	39	116	20	0.000	Bersaglieri et al 2004
She	China	39	116	20	0.000	Bersaglieri et al 2004
Naxi	China	39	116	20	0.000	Bersaglieri et al 2004
Tu	China	39	116	20	0.000	Bersaglieri et al 2004
Yakut	Siberia	65	125	50	0.060	Bersaglieri et al 2004
Japanese	Japan	35	139	62	0.000	Bersaglieri et al 2004
Papuan	New Guinea	-9	147	34	0.000	Bersaglieri et al 2004
Melanesian (NAN)	Bougainville	-9	159	44	0.000	Bersaglieri et al 2004
Yoruba	Nigeria	7	3	50	0.000	Bersaglieri et al 2004
Mozabite	Algeria	36	3	60	0.217	Bersaglieri et al 2004
San	Namibia	-22	17	14	0.000	Bersaglieri et al 2004
Bantu	South Africa	-26	28	16	0.000	Bersaglieri et al 2004
Bantu	N.E. Kenya	-1	36	24	0.000	Bersaglieri et al 2004
Pima	Mexican	19	-99	50	0.000	Bersaglieri et al 2004
Maya	Mexico	19	-99	50	0.020	Bersaglieri et al 2004
Colombian	Colombia	4	-74	26	0.000	Bersaglieri et al 2004
Karitiana	Brazil	-15	-47	48	0.000	Bersaglieri et al 2004
Surui	Brazil	-15	-47	42	0.000	Bersaglieri et al 2004
Greece	Greeks	38	24	82	0.134	TCGA
Turkey	Anatolian Turks	38	30	98	0.031	TCGA
Israel	Israeli Arabs	32	35	36	0.000	TCGA
Palestine	Palestinian Arabs	32	35	34	0.029	TCGA
Saudi Arabia	Saudi Bedouin	32	35	86	0.000	TCGA
Jordan	Jordanian Bedouin	32	36	40	0.075	TCGA
Ukraine	Ukrainian	48	36	92	0.217	TCGA
Syria	Assyrians	34	36	80	0.038	TCGA
Israel	Israeli Bedouin	33	37	26	0.000	TCGA
Azerbaijan	Azerbaijani	40	50	44	0.023	TCGA
Iran	Iranian	36	52	90	0.044	TCGA
Uzbekistan	Uzbekistani	40	64	36	0.000	TCGA
Afghanistan	Uzbek	35	68	76	0.079	TCGA
Afghanistan	Tadjik	36	69	98	0.102	TCGA

Afghanistan	Pashtu (Pushtu)	35	72	16	0.125	TCGA

B3 – GLOBAL DISTRIBUTION OF LACTASE PERSISTENCE FREQUENCIES

COUNTRY	POPULATION SUB-GROUP	REFERENCE	FREQ OF DIGESTORS	Latitude	Longitude
SOUTH AFRICA	BANTU	Jersky et al	0.048	-26.200	28.083
SOUTH AFRICA	Shanigaan	Segal et al	0.140	-26.200	28.083
SOUTH AFRICA	Sotho	Segal et al	0.348	-26.200	28.083
SOUTH AFRICA	Swazi	Segal et al	0.250	-26.200	28.083
SOUTH AFRICA	Tswana	Segal et al	0.167	-26.200	28.083
SOUTH AFRICA	Xhosa	Segal et al	0.176	-26.200	28.083
SOUTH AFRICA	Zulu	Segal et al	0.187	-26.200	28.083
BOTSWANA	HUA	Nurse et al	0.080	-24.983	25.350
KALAHARI	!KUNG	Jenkins et al	0.025	-24.033	21.900
TANZANIA	Bantu	Jackson et al	0.080	-6.833	36.983
TANZANIA	Masai	Jackson et al	0.380	-6.833	36.983
GABON	Bantu	Gendrel et al	0.364	0.383	9.450
CAMEROON	Fulani - nomadic	Kretchmer et al 1971	0.780	6.467	11.550
CAMEROON	Fulani - town	Kretchmer et al 1972	0.292	6.467	11.550
CAMEROON	Hausa	Kretchmer et al 1972	0.235	6.467	11.550
SUDAN	Nuba	Bayoumi et al	0.207	6.800	29.683
SUDAN	Dinka	Bayoumi et al	0.250	6.800	29.683
SUDAN	Nuer	Bayoumi et al	0.217	6.800	29.683
SUDAN	Shilluk	Bayoumi et al	0.279	6.800	29.683
SUDAN	Nilotic	Bayoumi et al	0.333	7.767	27.667
NIGERIA	Ibo	Olatunbosun et al	0.180	10.833	7.667
NIGERIA	Yoruba	Olatunbosun et al	0.160	10.833	7.667
SUDAN	Dongolawi	Bayoumi et al	0.187	11.750	26.933
SUDAN	Nubians	Bayoumi et al	0.333	11.750	26.933
SUDAN	Shaygi	Bayoumi et al	0.381	11.750	26.933
SUDAN	Habbani	Bayoumi et al	0.474	13.083	30.350
SUDAN	Misseri	Bayoumi et al	0.400	13.083	30.350
NIGER	Tuareg	Flatz et al	0.837	13.667	1.783
SENEGAL	Diolas	Arnold et al	0.725	14.667	-17.433
SENEGAL	Sereres	Arnold et al	0.711	14.667	-17.433
SENEGAL	Toucouleurs	Arnold et al	0.900	14.667	-17.433
SENEGAL	Wolof	Arnold et al	0.510	14.667	-17.433
SENEGAL	Peuhls	Arnold et al	1.000	15.400	-15.117
SUDAN	Jaali	Bayoumi et al	0.531	15.588	32.534
SUDAN	Amarar	Bayoumi et al	0.866	19.615	37.217
SUDAN	Artega	Bayoumi et al	0.818	19.615	37.217
SUDAN	Beni Amir	Bayoumi et al	0.875	19.615	37.217
SUDAN	Bisharin	Bayoumi et al	0.864	19.615	37.217
SUDAN	Haddendoa	Bayoumi et al	0.796	19.615	37.217
SUDAN	Bedja	Bayoumi et al	0.889	19.633	30.417
SUDAN	Gomoeia	Bayoumi et al	0.677	19.633	30.417
SUDAN	Kahli	Bayoumi et al	0.619	19.633	30.417
EGYPT	Upper Egypt, South	Hussein et al	0.400	24.088	32.899
EGYPT	Upper Egypt, North	Hussein et al	0.153	25.683	32.650
EGYPT	Suez Canal Zone	Hussein et al	0.312	29.791	32.550

EGYPT	Cairo and Giza	Hussein et al	0.328	30.050	31.250
EGYPT	Nile Delta	Hussein et al	0.268	30.791	30.998
TUNISIA	tunisien	Filali et al	0.204	36.803	10.180
AMERICA	mexican descent	Dill et al	0.455	9.400	-99.050
PUERTO RICO	Puerto Ricans	Goldman and Corcino	0.460	18.013	-66.614
MEXICO	Huamantha	Lisker et al	0.187	19.317	-97.933
MEXICO	Ixtenco	Lisker et al	0.237	19.317	-97.933
AMERICA	American Indians from Oklahoma	Bose and Welsh	0.194	33.334	-90.168
AMERICA	Oklahoma native american Indians	Caskey et al	0.133	33.334	-90.168
SINGAPORE	Singaporean	Bolin et al	0.000	1.283	103.838
SINGAPORE	Chinese	Yap et al	0.052	1.293	103.856
SRI LANKA	patients at the university of Ceylon	Senewiratne et al	0.275	6.932	79.848
INDIA	Pondicherry area	Tandon et al	0.550	8.483	76.917
THAILAND	North Thailand	Rotthauwe et al	0.000	13.750	100.517
THAILAND	Northern Thailand	Flatz et al	0.000	13.750	100.517
THAILAND	Thai	Keusch et al	0.029	13.750	100.517
Vietnam	Living in the USA	Nong The Anh et al	0.000	20.417	106.167
TAIWAN	Chinese	Juei-Low Sung and Ping Ling Shih	0.880	25.017	121.450
China	Northern region of Xinjiang	Yongfa et al	0.236	25.808	106.075
INDIA	Northern Indians	Gupta et al	0.729	28.600	77.200
INDIA	Trivandrum area	Tandon et al	0.600	28.600	77.200
INDIA	Delhi area	Tandon et al	0.726	28.667	77.217
Pakistan	Baloochi	S M Rab and A Baseer	1.000	32.271	71.920
Pakistan	Baluchistani	M Ahmad and G Flatz	0.375	32.271	71.920
Pakistan	Kashmiri	M Ahmad and G Flatz	0.296	32.271	71.920
Pakistan	Mohajir	S M Rab and A Baseer	0.800	32.271	71.920
Pakistan	Punjabi	M Ahmad and G Flatz	0.410	32.271	71.920
Pakistan	Punjabi	S M Rab and A Baseer	1.000	32.271	71.920
Pakistan	Sindhi	M Ahmad and G Flatz	0.424	32.271	71.920
Pakistan	Sindhi	S M Rab and A Baseer	1.000	32.271	71.920
Pakistan	Pathan	S M Rab and A Baseer	1.000	35.000	72.000
China	Inner Mongolia autonomous region	Yongfa et al	0.121	39.100	100.276
China	Northern Han	Yongfa et al	0.077	39.900	116.413
Japan	Japanese	Yoshida et al	0.275	40.583	140.467
RUSSIA	Khants	Lember et al	0.063	61.250	73.417
AUSTRALIA	Kimberly region, North `west oz	Brand et al	0.156	-23.700	133.883
PAPUA NEW GUINEA	Goodenough Island area	Gibney et al	0.172	-10.183	148.700
PAPUA NEW GUINEA	non-Austronesian	GC Cook	0.020	-9.483	147.183
PAPUA NEW GUINEA	Huki, Mendi and Dunai	Jenkins	0.100	-5.700	142.950
GREECE	Greek Cypriots	Kanaghinis et al	0.340	35.167	33.367
GREECE	Cretans	Kanaghinis et al	0.440	35.335	25.133
GREECE	Greeks	Spanidou and Petrakis	0.625	37.983	23.733
GREECE	Greeks	Kanaghinis et al	0.553	37.983	23.733
GREECE	Greeks	Zografos et al	0.768	37.983	23.733
ITALY	Sicilian	Burgio et al	0.290	38.117	13.367
SPAIN	Galicia	Leis et al	0.650	40.400	-3.683
ITALY	Neapolitan region	Ritis et al	0.000	40.833	14.250
FRANCE	Western France	Cloarec et al	0.765	44.933	5.817
ITALY	Italians	Bozzani et al	0.381	45.467	9.200
ITALY	Lombard, Pietmonte, veneto	Burgio et al	0.490	45.699	12.226
HUNGARY	East	Czeizel et al	0.714	46.250	20.167

HUNGARY	West	Czeizel et al	0.720	46.583	17.417
AUSTRIA	karnten region	Rosencranz et al	0.804	46.983	15.900
AUSTRIA	Steiermark region	Rosencranz et al	0.754	46.983	15.900
HUNGARY	Mixed	Czeizel et al	0.592	47.500	19.083
AUSTRIA	Tirol, Vorarlberg region	Rosencranz et al	0.831	47.700	15.550
GERMANY	Bayern region	Flatz et al	0.852	47.800	12.533
AUSTRIA	Oberosterreich, Salzburg	Rosencranz et al	0.844	47.800	13.033
HUNGARY	Matyo	Czeizel et al	0.366	47.817	20.583
HUNGARY	Romai	Czeizel et al	0.558	47.950	21.717
HUNGARY	Northeast	Czeizel et al	0.583	48.100	20.783
AUSTRIA	4 grandparents from Austria	Rosencranz et al	0.795	48.200	16.367
GERMANY	Baden-Wurttemberg region	Flatz et al	0.781	48.800	9.200
FRANCE	North French	Cuddenec et al	0.774	48.867	2.333
GERMANY	Rheinland-Pfalz and Saarland region	Flatz et al	0.911	49.517	6.617
CANADA	Czech Slavs	Leichter	0.824	50.083	14.467
POLAND	South	Socha et al	0.621	50.083	19.917
CZECH REPUBLIC	Living in district of Plzan	Madzariviva-Nohejlova et al	0.150	50.417	14.967
GERMANY	Nordrhein-Westfalen region	Flatz et al	0.873	51.317	7.150
BRITAIN	British natives	Ho et al	0.947	51.754	-1.254
BRITAIN	British natives, Oxford area	Pena et al	0.900	51.754	-1.254
GERMANY	Hessen region	Flatz et al	0.871	52.017	10.783
POLAND	central	Socha et al	0.630	52.233	19.367
CANADA	Polish slavs	Leichter	0.714	52.250	21.000
POLAND	Mixed	Socha et al	0.635	52.250	21.000
BRITAIN	White	Iqbal et al	0.970	52.485	-1.860
GERMANY	Berlin, Foreign and Unknown	Flatz et al	0.933	52.517	13.400
POLAND	East	Socha et al	0.629	52.583	17.867
GERMANY	Niedersachsen and Bremen region	Flatz et al	0.857	53.083	8.800
GERMANY	Hanover	Howell et al	0.867	53.183	8.517
IRELAND	Irish	Fielding et al	0.960	53.333	-6.250
GERMANY	Schkeswig-Holstein and Hamburg region	Flatz et al	0.875	53.550	10.000
POLAND	North-East	Socha et al	0.588	53.833	22.350
RUSSIA	Komi-Izhems	Kozlov	0.375	55.000	82.717
RUSSIA	Khanty (Northern)	Kozlov	0.287	55.750	37.583
RUSSIA	Kildin Saami	Kozlov	0.520	55.750	37.583
RUSSIA	Mansi	Kozlov	0.284	55.750	37.583
RUSSIA	Komi-Permiaks	Kozlov	0.500	55.750	37.583
RUSSIA	Udmurtians	Kozlov	0.413	55.750	37.583
RUSSIA	Udmurtian rep.	Kozlov	0.600	55.750	37.583
RUSSIA	West-Siberian	Kozlov	0.511	55.750	37.583
BRITAIN	White	FERGUSON et al	0.953	55.949	-3.161
ESTONIA	Estonians	Lember et al	0.750	59.434	24.728
ESTONIA	Setus	Lember et al	0.400	59.434	24.728
FINLAND	Rural Finns	Jussila et al	0.824	60.467	25.383
FINLAND	Finnish-speaking	Sahi	0.827	60.600	21.433
FINLAND	Swedish-speaking	Sahi	0.923	60.600	21.433
RUSSIA	Nenets (West Siberia)	Kozlov	0.222	61.250	73.417
KUWAIT	(Arab) Kuwaiti	Sanae et al	0.529	29.370	47.978
AFGHANISTAN	Pashtun	Rahimi et al	0.211	31.617	65.717
CANADA	Jews living in Western Canada	Joseph Leichter	0.313	31.767	35.233
JORDAN	Arabs from the urban and agricultural zone	Hijazi et al	0.250	31.950	35.933
JORDAN	Jordanian Bedouin	Hijazi et al	0.759	31.950	35.933
JORDAN	Arabians of the	Snook et al	0.232	31.950	35.933

	Mediterranean basin				
LEBANON	Lebanese	Nasrallah	0.227	33.872	35.510
AFGHANISTAN	Pasha-I	Rahimi et al	0.133	34.517	69.183
AFGHANISTAN	Mixed urban	Rahimi et al	0.235	34.517	69.183
AFGHANISTAN	Hazara	Rahimi et al	0.200	35.000	66.000
IRAN	Iranians, nr Tehran	Sadre et al	0.175	35.672	51.424
AFGHANISTAN	Uzbek	Rahimi et al	0.000	36.183	68.733
TURKEY	South Coast of Turkey	Flatz et al	0.278	37.017	35.300
TURKEY	Eastern Anatolia	Flatz et al	0.262	38.500	43.383
TURKEY	Western Anatolia and European Turkey	Flatz et al	0.302	38.683	29.417
TURKEY	Central Anatolia	Flatz et al	0.288	40.683	31.667
TURKEY	North Coast of Turkey	Flatz et al	0.313	41.567	35.933

Appendix C

The Causal Element for the Lactase Persistence/ non-persistence Polymorphism is Located in a 1 Mb Region of Linkage Disequilibrium in Europeans

M. Poulter[1*], E. Hollox[1**], C. B. Harvey[1***], C. Mulcare[1], K. Peuhkuri[2,3], K. Kajander[2,4], M. Sarner[5], R. Korpela[2,3,4] and D. M. Swallow[1]

[1]*The Galton Laboratory, Department of Biology, Wolfson House, University College London, 4 Stephenson Way, London NW1 2HE, UK*

[2]*Valio Ltd PO Box 30 FIN-00039 Valio Helsinki, Finland*

[3]*Foundation for Nutrition Research, Helsinki, Finland*

[4]*Institute of Biomedicine, Department of Pharmacology and Toxicology, University of Helsinki, Finland*

[5]*Department of Gastroenterology, University College London Hospitals, London WC1N 8AA, UK*

Summary

Expression of lactase in the intestine persists into adult life in some people and not others, and this is due to a *cis*-acting regulatory polymorphism. Previous data indicated that a mutation leading to lactase persistence had occurred on the background of a 60 kb 11-site *LCT* haplotype known as **A** (Hollox *et al.* 2001). Recent studies reported a 100% correlation of lactase persistence with the presence of the T allele at a CT SNP at −14 kb from *LCT*, in individuals of Finnish origin, suggesting that this SNP may be causal of the lactase persistence polymorphism, and also reported a very tight association with a second SNP (GA −22 kb) (Enattah *et al.* 2002). Here we report the existence of a one megabase stretch of linkage disequilibrium in the region of *LCT* and show that the −14 kb T allele and the −22 kb A allele both occur on the background of a very extended **A** haplotype. In a series of Finnish individuals we found a strong correlation (40/41 people) with lactose digestion and the presence of the T allele. The T allele was present in all 36 lactase persistent individuals from the UK (phenotyped by enzyme assay) studied, 31/36 of whom were of Northern European ancestry, but not in 11 non-persistent individuals who were mainly of non-UK ancestry. However, the CT heterozygotes did not show intermediate lactase enzyme activity, unlike those previously phenotyped by determining allelic transcript expression. Furthermore the one lactase persistent homozygote identified by having equally high expression of **A** and **B** haplotype transcripts, was heterozygous for CT at the −14 kb site. SNP analysis across the 1 megabase region in this person showed no evidence of recombination on either chromosome between the −14 kb SNP and *LCT*. The combined data shows that although the −14 kb CT SNP is an excellent candidate for the cause of the lactase persistence polymorphism, linkage disequilibrium extends far beyond the region searched so far. In addition, the CT SNP does not, on its own, explain all the variation in expression of *LCT*, suggesting the possibility of genetic heterogeneity.

*Current addresses: MRC Prion Unit, University College London Institute of Neurology, Queen Square House, Queen Square, London WC1N 3BG, UK

**Institute of Genetics, School of Medicine, Queen's Medical Centre, University of Nottingham, Nottingham NG7 2UH, UK

***Molecular Endocrinology Group, Faculty of Medicine, Imperial College London, Clinical Research Building, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK

Introduction

Intestinal lactase activity persists into adult life in some people but not others, and the molecular basis of this genetic trait is not understood (Swallow & Hollox, 2000). Lactase persistence is a monogenic trait which is

inherited in an autosomal dominant manner. Monoallelic expression of lactase transcripts in heterozygotes (Wang *et al.* 1998, 1995), and more recently family studies (Enattah *et al.* 2002), demonstrated clearly that the polymorphism was controlled by a *cis*-acting mechanism – i.e. the causative polymorphism was in a cis-acting regulatory element near the lactase gene. This was consistent with early studies that reported a trimodal distribution of lactase activity in jejunal samples from unrelated UK or German adults (Flatz 1984; Ho *et al.* 1982). Our population genetic data indicated that a mutation leading to lactase persistence had occurred on the background of a particular haplotype of the gene (Harvey *et al.* 1998; Hollox *et al.* 2001). This 60 kb 11-site haplotype (**A**) covering the lactase gene (*LCT*) is extremely common in Northern Europe where lactase persistence is also common, and lactase persistence is associated with the **A** haplotype (Harvey *et al.* 1998), although occasionally lactase persistence was found in combination with a non-**A** haplotype (Harvey *et al.* 1998; Wang *et al.* 1995). A recent study reported a complete association of lactase persistence with the presence of the T allele at a CT SNP at −14 kb from *LCT*, and suggested that this SNP may be causal of the lactase persistence polymorphism (Enattah *et al.* 2002). A second SNP at −22 kb was also highly associated. The study was conducted mainly in Finnish individuals, and only 5 samples from persistent non-Finnish individuals were tested. The aims of our study were to determine: the extent of linkage disequilibrium upstream from *LCT*; whether the T allele at −14 kb and the A allele at −22 kb were mutations uniquely on the background of the **A** haplotype; and whether 100% correlation of the −14 kb SNP could be found in our series of samples.

Methods

Contig Construction

A contig centred on *LCT* was constructed by chromosome walking using three high-density gridded libraries distributed via the UK HGMP Resource Centre. These were the chromosome 2 specific Lawrence Livermore National Laboratory fosmid library LL02NC03 (average insert size 35–40 Kb) and a cosmid library LL02NC02 (average insert size 40 kb) and the PAC library LL02NP04 (85 Kb), and the whole genome PAC library RPCI1 (110 Kb) produced by the Roswell Park Cancer Institute. Single copy probes from *D2S442* and *LCT* were used to screen the libraries and single copy sequence generated from both ends of positive clones. This sequence was used to generate single-copy PCR products, to rescreen the libraries, and chromosome walking was continued using this approach. Mapping the ends of *D2S442* positive clones on to previously mapped YACs (Jarvela *et al.* 1998) oriented them with respect to *LCT*, and further orientations were deduced by the patterns of positive and negative signal obtained by PCR amplification of single copy sequence using other clones as templates. BAC clones from the RPCI-11 library, arranged in contigs and sequenced to first draft standard at the Genome Sequencing Center at the University of Washington Medical School at St Louis, were used to extend the map and particularly to identify SNPs on the other side of *LCT*.

All hybridisations using ^{32}P- labelled probes were performed in 6×SSC (20×SSC = 3M NaCl, 0.3M sodium citrate @ pH7.0, 1×Denhardts solution (100 × = 2%[w/v] Ficoll, 2%[w/v] Polyvinylpyrrolidone and 2%[w/v] Bovine serum albumin @ pH 7.2) and 50μg/ml of sonicated herring sperm DNA at 65°C. The filter membranes were washed to a final stringency of 0.2×SSC 0.1% SDS at 65°C and subjected to autoradiography using standard procedures.

Clone DNA was isolated with Qiagen Plasmid Kits (Qiagen) using the very low copy protocol. PCR was performed on either 1μl of boiled bacterial culture or 100ng of genomic DNA with 25 picomoles of each primer, 0.2 mM dNTPs, 75 mM Tris-HCl, pH 9.0 at 25°C, 20 mM $(NH_4)_2SO_4$, 1.5 mM $MgCl_2$, 0.1%[w/v] Tween and 1.25 U *Taq* Polymerase (ABgen) annealing at 57-59°C for 30 sec and extension at 72°C for 1 min for 32 cycles. PCR products were analysed on 2% agarose and detected by ethidium bromide staining. Direct double-stranded sequencing of clone DNA (~2μg) with vector primers used the Thermo Sequenase Radiolabelled Terminator Cycle Sequencing Kit (Amersham).

As sequences were obtained they were analysed for the presence of repeat motifs and unique sequences on the databases through the GCG Wisconsin Package of programs at the HGMP Resource Centre, Hinxton Hall

Cambridge, UK and programs available through the NCBI website (http://www.ncbi.nlm.nih.gov/).

Family Samples

The French families from the CEPH series (Dausset *et al.* 1990) were studied in detail. The maximum possible number of extended haplotypes was generated from each family by analysis of parents and selected children and/or grandparents. A few other CEPH samples were used in order to further characterise extended C haplotype chromosomes, which were rare in the French population. These were from individuals 1334: 10, 11, 12 and 13, 1424: 11 and 12; and 1447: 9 and 10.

Phenotyped Individuals

We lactose tolerance tested 42 unrelated individuals who were all healthy, by giving a 50 g lactose load after an overnight fast, and using three methods for determining tolerance. These were breath hydrogen, urinary galactose and blood glucose (Peuhkuri *et al.* 1998). The cut-offs described previously were used (Peuhkuri *et al.* 1998). The "gold standard procedure" (all three measures taken, and diagnosis made on the basis of two or more concordant results) (Peuhkuri *et al.* 1998) was used in most cases. Breath hydrogen alone was measured in 13 of the earlier cases. Details of symptoms were also recorded. In one case 2/3 results were borderline so this individual was excluded from the study. Ethical approval for this study was obtained from the Joint Authority for the Hospital District of Helsinki and Uusimaa (HUS) Ethics Committee.

A cohort of 48 UK adults, collected in University College Hospitals London (with ethics approval from UCL/UCLH Committee on the Ethics of Clinical Investigations) and used in our previous lactase studies, was also tested (Harvey *et al.* 1995a; Wang *et al.* 1995). It should be noted that only cases showing normal villous architecture and immunohistology for sucrase-isomaltase and alkaline phosphatase were included in this cohort. The lactase persistence status of these individuals was determined directly by assay of lactase and sucrase and measuring the sucrase/lactase (S/L) ratio. Eleven of the individuals were diagnosed as non-persistent (S/L ratio of more than 10) and one individual gave an ambiguous result (ratio of 7.7). 18/48 individuals were heterozygous for exonic SNPs (9 lactase non-persistent and 9 lactase persistent) and the relative level of expression of the lactase transcripts could thus be determined. 8/9 of these informative lactase persistent individuals were diagnosed as heterozygous for lactase persistence, since they showed high expression of one allele and low expression of the other. In one case both transcripts were expressed at equal levels, which was considered diagnostic of homozygosity for lactase persistence. The remainder showed low expression of both transcripts, and were lactase non-persistent. Five of the individuals (three persistent and two non-persistent), who are heterozygous for exonic polymorphisms (and thus of known lactase persistence genotype), were selected from this cohort and constituted the panel of samples for sequencing and SNP searching. Genomic DNA used was from lymphoblastoid cell lines derived from the peripheral blood of these individuals and/or DNA extracted from the biopsies. Critical results were confirmed on biopsy DNA.

SNP Identification and Typing

PCR products corresponding to clone end sequences were generated from genomic DNA from the panel of the five individuals of known lactase persistence genotype, and directly sequenced to identify SNPs. These, and two recently published SNPs at −14 kb and −22 kb (Enattah *et al.* 2002), were typed on further samples using a variety of methods: Restriction Fragment Length Polymorphism (RFLP) analysis, RFLP generated by primer design (Thomas *et al.* 1999), Tetra ARMS PCR (Ye *et al.* 2001); and direct sequencing of PCR products. The approximate location of the SNPs was determined using average clone size information and electronically published but unfinished sequence data. The sites of the SNPs in relation to the latest Golden Path sequence, as found on the June 2002 freeze of the Draft Human Genome Browser (http://genome.cse.ucsc.edu/), and the flanking sequences can be found on our web site (http://www.gene.ucl.ac.uk/mucin/). Throughout this study we have interpreted the observed phenotypes at each locus as genotypes, making the assumption that there are no silent/deleted alleles; hence C, CT and T are written and interpreted as CC, CT and TT respectively. The insertion/deletion polymorphism in

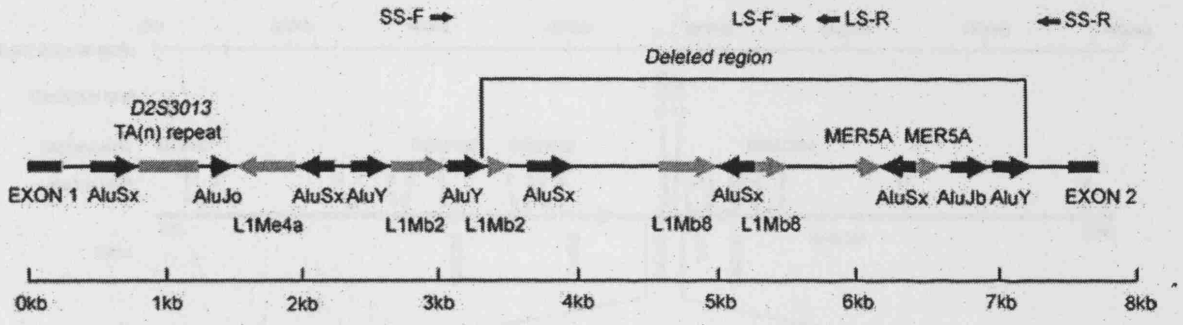


Figure 1 Diagram schematically showing the allelic difference due to the INDEL polymorphism in intron 1. The primer positions used for testing for the presence of the S and L alleles are shown. Note the abundance of repeat elements, and that the long allele contains an extra 3.5 kb of sequence which is surrounded by two AluY elements, only one of which is present in the short allele. The deleted segment is indicated. SS-F = Short intron 1 specific forward primer; SS-R = Short intron 1 specific reverse primer; LS-F = Long intron 1 specific forward primer; LS-R = Long intron 1 specific reverse primer

intron 1 (*INDEL* intron 1) was typed using two separate PCRs, one for each allele (Figure 1), each reaction including control PCRs of similar size to check for DNA and PCR quality. The primers used for the long allele PCR were: 5′GTGGAATGTGAAACGGATCC3′ (LS-F) and 5′AGGACCATATGGCTGTCTTC3′ (LS-R), product size 244bp, and were both located in the deleted sequence. For the short allele PCR the primers were: 5′CTAGGACATCATAGCTGCCT3′ (SS-F), and 5′CTCTGACTGTGGAAACCACTG3′ (SS-R), product size 944 bp. The longer product size used for the short allele PCR was needed to avoid repeat elements. For both PCRs the conditions were: initial denaturation at 95°C for 5 minutes, followed by 32 cycles of denaturation at 95°C for 30 seconds, annealing at 59°C for 30 seconds, and extension at 72°C for 2 minutes for the short allele PCR, and for 1 min. for the long allele PCR. Details of the primers for the control products included in the PCR are given on our website (http://www.gene.ucl.ac.uk/mucin/).

Haplotype Determination

Haplotypes across the whole region were deduced by segregation, from the CEPH family data in a total of 64 chromosomes by testing sufficient parents and/or children. For the unrelated Finnish chromosomes haplotypes were deduced by using information obtained from the homozygotes, comparison with the CEPH haplotypes, and assuming the minimum number of historic recombination events.

Linkage Disequilibrium Analysis

Pairwise linkage disequilibrium between combinations of the loci was estimated using the 48 CEPH chromosomes of French origin. Linkage disequilibrium was measured using the normalised D coefficient D′, which takes values between 0 for complete equilibrium to 1 for complete disequilibrium. D' was calculated using $D' = |D_{ij}/D_{max}|$ where $D_{ij} = p_{ij}-p_ip_j$ and $D_{max} = \min(p_ip_j, (1-p_i)(1-p_j))$ if $D_{ij} < 0$ or $\min((1-p_i)p_j, p_i(1-p_j))$ if $D_{ij} > 0$. p_i and p_j are the frequencies of alleles i and j, and p_{ij} is the frequency of the haplotype having i at the first locus and j at the second locus, using the computer program HaploXT (http://www.sph.umich.edu/csg/abecasis/GOLD/docs/haploxt.html). Haplotypes of the unrelated Finnish individuals were deduced by consideration of the allele combinations found in the homozygous individuals and in the CEPH samples, and assuming the minimum number of ancestral recombinations.

Association Analysis

Pairwise association of alleles at each of the SNP loci with lactose digestion status was analysed by Fisher's exact test (1 sided p-values).

Other Statistical Tests

The distribution of activities of the individuals of different genotype was compared using a Students T-test and the Mann-Whitney Rank Test, and the distribution of the activity ratios using the Mann-Whitney Rank Test.

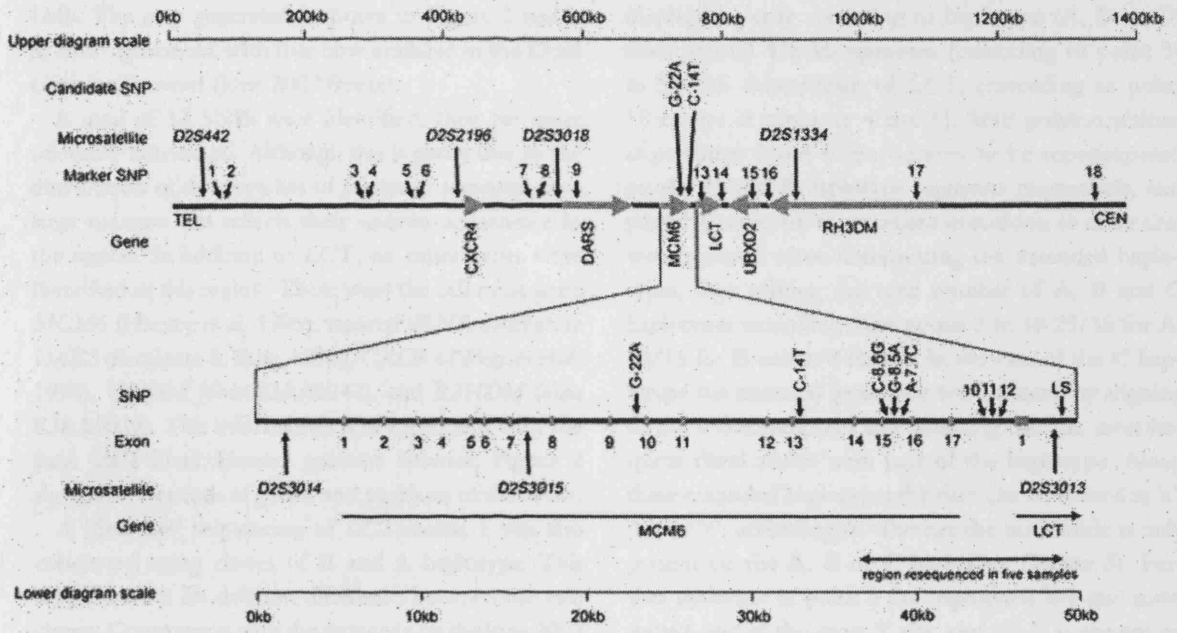


Figure 2 Map of a 1.2 megabase region surrounding *LCT*. The SNPs that were tested in this study are shown with arrows, and the ones used in the initial haplotype construction are numbered. The enlarged section shows the region sequenced by Enattah and colleagues (Enattah *et al.* 2002); the region also completely sequenced right through in five people of known lactase persistence genotype in this project is also shown.

Determination of Allelic Expression

Relative mRNA transcript levels were determined using PCR products obtained by RT-PCR performed under the general conditions described previously (Wang *et al.* 1994) with appropriate controls and 'no RT' blanks. The primers for the RT-PCR are located in exons 1 and 3 of *LCT*, and 30 cycles of amplification was selected to produce sufficient template DNA for sequencing. The transcripts were distinguished using exonic SNPs at nt 593 and nt 666 and quantification by phosphorimage analysis of ^{33}P sequencing gels, as described previously (Wang *et al.* 1998).

Results

Construction of Extended Lactase Haplotypes and Association Analyses

Sequence 5′ and 3′ of the lactase gene:

Initially 10 kb 5′ of exon 1 and 2.7 kb 3′ of exon 17 was sequenced using existing clones (Boll *et al.* 1991). Overlapping primer sets were designed on this sequence and used to produce amplicons from the panel of five samples with known lactase persistence genotypes. Nine new polymorphic sites were detected upstream of the sites we had already reported (Wang *et al.* 1998). In the five samples all of the polymorphisms were in complete association with alleles at one or other of the three previously published polymorphic sites that define the three major haplotypes (Harvey *et al.* 1995a). The new sites found are in agreement with those recently reported (Enattah *et al.* 2002), except that we found an additional site at approximately −7.7 kb. Nine new polymorphic sites were found at the 3′ side of the gene (7 SNPs and 2 microsatellites). Eight of these were also associated with the *LCT* haplotypes, while one SNP showed a novel allele in panel sample 4. Further analysis of this variant showed that it was unique to this individual.

Contig Construction and Distribution of Polymorphisms and Coding Sequences

A contig of 60 overlapping BAC, PAC fosmid, and cosmid clones was constructed that spanned approximately

1Mb. The map generated is shown in Figure 2 and is in close agreement with that now available in the Draft Genome Browser (June 2002 freeze).

A total of 13 SNPs were identified; they are quite unevenly distributed. Although this is partly due to the distribution of the stretches of sequence screened, to a large measure this reflects their uneven occurrence in the region. In addition to *LCT*, six other genes were identified in this region. These were the cell cycle gene *MCM6* (Harvey *et al.* 1996), aspartyl tRNA synthetase *DARS* (Escalante & Yang, 1993), *CXCR4* (Wegner *et al.* 1998), *UBXD2* (alias KIAA0242), and *R3HDM* (alias KIAA0029). This information is in agreement with the June 2002 Draft Human genome browser. Figure 2 shows the locations of genes, and positions of the SNPs.

A 'first pass' sequencing of *LCT* intron 1 was also conducted using clones of **B** and **A** haplotype. This revealed a 3.5 kb deletion difference between the two clones. Comparison with the sequence on the June 2002 Draft Human Genome Browser shows that this additional sequence inserts at position 134245902, to give the *INDEL* intron 1 L allele. Primers were designed to test for this deletion (Figure 1).

Initial Haplotype Analysis

Extended haplotypes of 48 chromosomes were determined from the CEPH families from Northern France. The *LCT* haplotypes of these chromosomes have been reported previously (Harvey *et al.* 1995a) and comprise 28 **A**, 13 **B**, and 3 **C**, together with 2 **D**, 1 **E** and 1 **F**. For the purposes of the haplotype analyses; the two **D** chromosomes were considered as **B** chromosomes, since **D** is derived from **B** by a point mutation at nucleotide –875 (Harvey *et al.* 1995a; Hollox *et al.* 2001); and the chromosome **F** also as **B** since they differ only at the exon 17 SNP T5579C. To these chromosomes were added chromosomes from 8 individuals from the CEPH Utah families of Northern European ancestry, which added a further 5 **C** chromosomes, 10 **A** chromosomes and 1 **E**. The total number of chromosomes was therefore 64. For this part of the analysis 18 SNPs were tested, numbered 1–18 in Figure 2.

The majority of chromosomes showed the same few haplotypes previously seen over 60 kb, extended over very much longer distances. 36/64 chromosomes were identical by state according to haplotype (**A**, **B** or **C**) from around 420 kb upstream (extending to point 3) to 300 kb downstream of *LCT*, (extending to point 18 except at positions 4 and 6): SNP polymorphisms at positions 4 and 6 can be seen to be superimposed on the **B** and **A** haplotype segments respectively, and probably represent more recent mutations, so these sites were ignored when determining the extended haplotypes, thus making the total number of **A**, **B** and **C** haplotypes extending from points 3 to 18 25/38 for **A**, 10/16 for **B** and 1/8 for **C**). In the case of the **C** haplotype the ancestral haplotype was deduced by aligning all the 8 chromosomes and assuming that the most frequent distal alleles were part of the haplotype. Along these extended haplotypes the sites can be classed as 'a', 'b', or 'c', according to whether the nucleotide is only present on the **A**, **B** or **C** haplotype (Figure 3). Further upstream of point 3 the haplotypes become more mixed, and at the most 3′ site, site 18, C is present on all the **A** haplotype chromosomes while several of the **B** and **C** haplotype chromosomes are recombined at this point. The two **E** haplotype chromosomes support the proposed recombinant origin of **E** haplotype chromosomes (Hollox *et al.* 2001), the 5′ side representing the **C** haplotype and the 3′ side the **A** haplotype. Of the remaining 26 chromosomes, in 16 cases the pattern of alleles either side of the breakpoint was consistent with the idea that a simple recombination event has occurred which swapped the peripheral parts of the other common haplotypes (see our website for supplementary information).

The same full set of polymorphic sites was also tested in a series of 21 samples from unrelated individuals of Finnish ancestry and haplotypes were deduced from analysis of the homozygotes as well as comparison with the CEPH chromosomes, and assuming the minimum number of ancestral recombination events. 20/42 chromosomes showed the same extended haplotypes as in the CEPH individuals (11/23 **A**, 4/8 **B**, 5/10 **C**, IE).

Addition of Polymorphisms that Subdivide, and thus Post-Date, the A Haplotype

The insertion deletion polymorphism in intron 1 (*INDEL intron 1)* was also analysed in the CEPH families, and full haplotype analysis of this and the newly

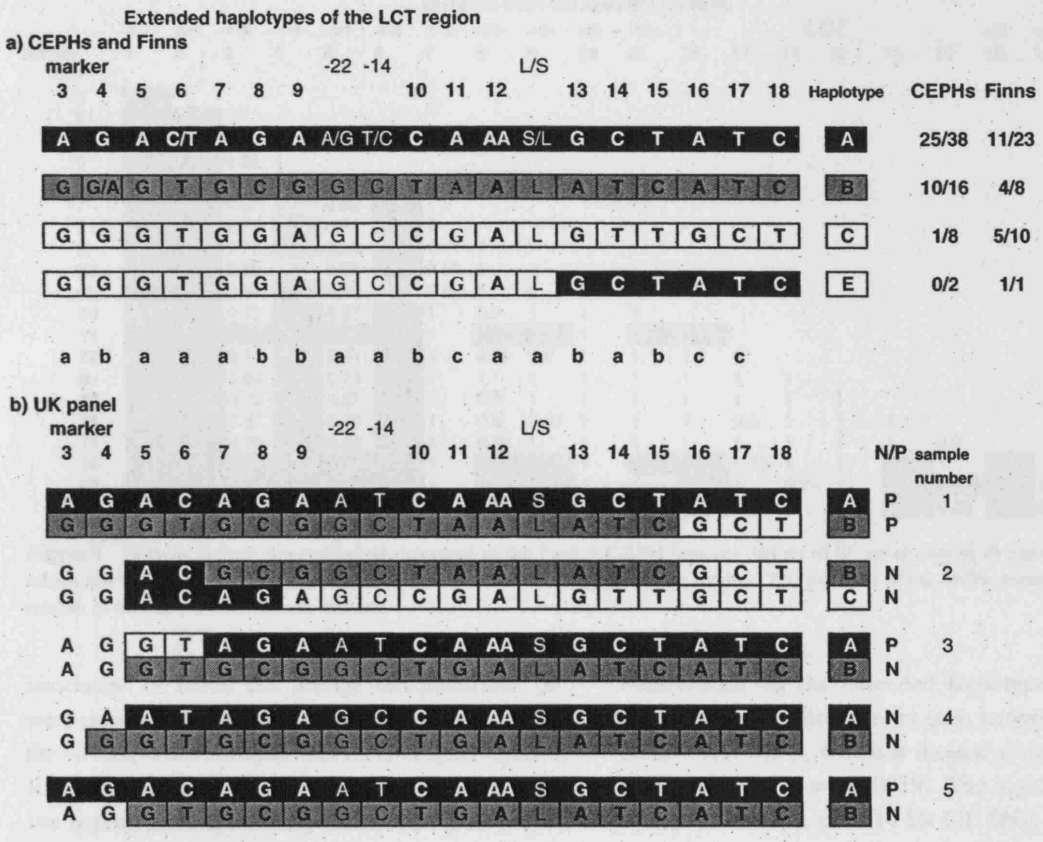


Figure 3 Extended haplotypes constructed using the original set of numbered markers, which are shown in bold. The new polymorphisms (not in bold) which were tested on all the French CEPH families (48 chromosomes), all the Finns, and all of the panel members 1-5, are superimposed. The **A** haplotype distinguished by markers 'a' is shown in black, the **B** haplotype with 'b' markers is shown in grey, and the **C** haplotype distinguished by 'c' markers is shown in white. A: Extended haplotypes found in the CEPH families and in the Finnish cohort. The number of chromosomes of each extended haplotype are indicated. B: Extended haplotypes found in the panel of 5 individuals whose lactase persistence genotype was deduced by comparison with the CEPH chromosomes.

described –14 kb CT and –22 kb GA SNPs (Enattah *et al.* 2002) was undertaken for all of the 48 French CEPH chromosomes.

INDEL intron 1: The L allele was carried by all **B** and **C** haplotype chromosomes, while most of the **A** haplotype chromosomes carried S. However, two of the 26 **A** chromosomes tested in this series carried an L allele.

CT-14 kb GA-22 kb: Similarly, all non-**A** chromosomes carried –14 kbC and –22 kbG, and all the –14 kb T and –22 kb A alleles were carried on **A** haplotype chromosomes. The two **A** haplotype chromosomes that carried the INDEL intron 1 L allele (AL) also carried –14 kbC and –22 kbG, but two other **A** haplotype chromosomes with intron 1 S (AS) also carried –14 kbC and –22 kbG. In two cases, –14 kbC was present together with –22 kbA on an 'AS' haplotype chromosome. Interestingly none of the **A** haplotype –14 kb C and –22 kb G chromosomes carried **A** haplotype markers over the full distance between marker 3 and 18, unlike the T and A carrying chromosomes in which 18/22 were on the background of full length **A** haplotype chromosomes.

Linkage Disequilibrium

The low level of haplotype diversity was reflected in the patterns of linkage disequilibrium. Pairwise

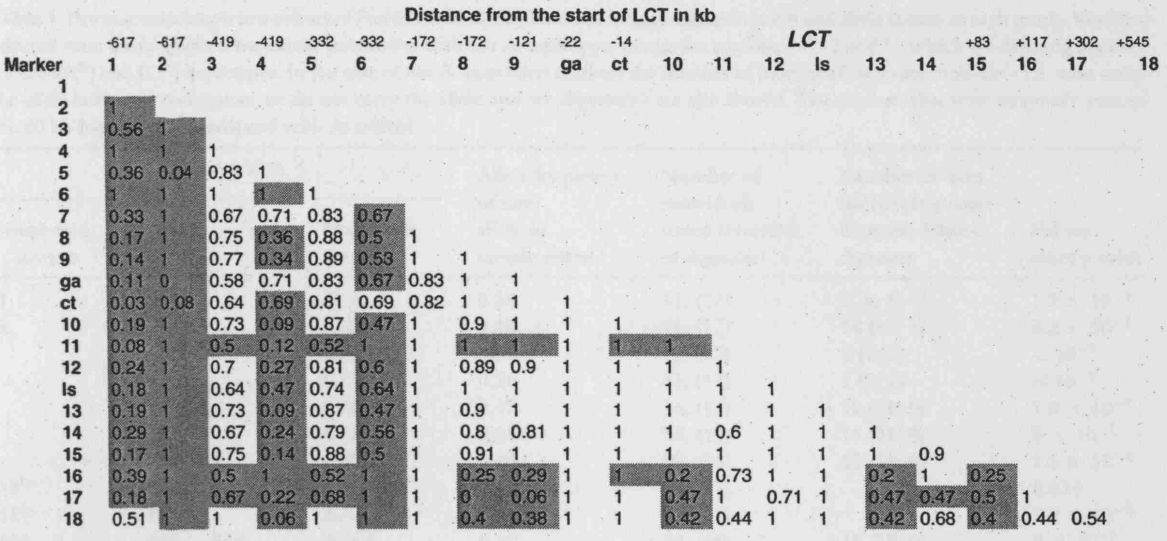


Figure 4 Pairwise linkage disequilibrium measured in the French CEPH families. Values of D′ are shown in all cases but those which not were significantly different from D′ = 0 at the 5% level or less are shaded. The position of the SNPs, measured in kb relative to the start of *LCT*, is also shown.

association of alleles and linkage disequilibrium (D′) were estimated between all of the informative sites, using the 48 CEPH chromosomes of French origin (Figure 4). It can be seen that there is significant LD across the entire region, there even being statistically significant LD ($p = 0.01$) between the first and last sites, which are separated by over one megabase. LD was not determined in the Finnish cohort since this population was selected artificially to contain more lactase non-persistent people than in the general population.

Allelic Association with Lactose Digestion (Lactase Persistence) Phenotype

For this part of the study we tested all 41 phenotyped Finnish individuals, but focussed on markers which discriminate **A** and non-**A** chromosomes, although all the haplotype defining markers within the gene were also tested. Three additional polymorphic sites located between –7.5 kb and –8.5 kb were also tested, since analysis of the original panel of five individuals had indicated that these also discriminate **A** and non-**A** chromosomes.

Each of the sites was assessed in relation to lactose digestion phenotype; phenotype was highly associated with several of the markers (Table 1). Consistent with observations on the extended haplotypes, statistically significant association can be seen between lactose digestion and site 3, which is located about 420 kb upstream of *LCT* ($p = 0.00019$). The highest association was however with the –14 kb CT SNP, in agreement with the recently reported results, but *one* individual gave a discrepant result, an apparent digester who does not carry a T at –14 kb (or a G at –22 kb). Examination of the detailed tolerance results from this individual showed conflicting results for one of the tests (blood glucose and urinary galactose both showed levels clearly in the 'digester' range, while the high breath hydrogen indicated maldigestion). The –22 kb marker was also very highly associated. In all but one case there was concordance between CT – 14 kb and GA-22 kb. This one discordant individual (–14 kb and –22 kb GA) was non-persistent (based on three tests as well as presence of symptoms).

Examination of the 82 putative Finnish haplotypes reveals that 12 of the 22 **A** haplotype chromosomes which carry the C allele at –14 kb show non **A** haplotype markers immediately upstream of –14 kb. In five cases non **A** haplotype markers were also found downstream of –14 kb, suggesting that the surrounding chromosomal fragment is of non A haplotype origin. Four of these five chromosomes carried the ancestral L allele

Table 1 Pairwise association in a cohort of Finnish individuals, between lactose digestion status and allele counts in each group. Markers selected were those which show strong association with the **A** haplotype, except for markers 10, 13 and 11 which are defining markers for the **B** (b) and **C** (c) haplotypes. In the case of the **A**-associated markers the number of individuals who are 'non-fits'- i.e. who carry the allele but are non-digesters, or do not carry the allele and are digesters - are also shown. The markers that were originally part of the 60 kb haplotype are indicated with an asterisk

Polymorphism			Allele frequency of rarer allele in sample tested	Number of individuals tested (Number of digesters)	Number of 'non fits'(Allele+/non digester; Allele-/ digester)	Fishers exact p value
Numbered marker	Position	Nucleotide change				
3	**−420 kb**	G/A	0.38	41, (17)	9 (8; 1)	1.9×10^{-4}
5	**−330 kb**	A/G	0.48	41, (17)	14 (13; 1)	8.2×10^{-3}
	−22 kb	G/A	0.27	41, (17)	2 (1; 1)	$< 10^{-8}$
	−14 kb	C/T	0.26	41, (17)	1 (0; 1)	$< 10^{-8}$
	−8.6 kb	C/G	0.47	36, (14)	11 (11; 0)	1.8×10^{-4}
	−8.5 kb	G/A	0.48	36, (14)	11 (11; 0)	5×10^{-5}
	−7.7 kb	A/C	0.49	39, (16)	13 (13; 0)	1.5×10^{-4}
10[b*]	**−958**	C/T	0.23	41, (17)		0.034
11[c*]	**−678**	A/G	0.23	41, (17)		8.2×10^{-3}
12*	**−552/−559**	A/AA	0.50	38, (15)	15 (15; 0)	8×10^{-5}
	Intron 1	IN/DEL L/S	0.49	41, (17)	12 (12; 0)	2×10^{-5}
13[b*]	**nt 666 (cDNA)**	G/A	0.23	41, (17)		0.034
14*	**nt 5579 (cDNA)**	C/T	0.44	41, (17)	16 (16; 0)	5×10^{-4}

in intron 1. In just 6/22 cases the C allele was found on a fully extended **A** haplotype chromosome. This contrasts with the finding that 19/21 T carrying chromosomes were on the fully extended haplotype background, and is consistent with the observation in the CEPH families of shorter blocks of **A** haplotype in −14 kbC carrying chromosomes, and longer blocks in −14 kbT carrying chromosomes.

Analysis of CT-14 kb and GA-22 kb in a Series of 48 Individuals Phenotyped and Genotyped for Lactase Persistence

In previous studies we characterised the level of lactase and lactase mRNA transcripts in duodenal biopsies. The level of lactase activity measured as a ratio of sucrase to lactase (S/L) showed the trimodal distribution that had originally been part of the evidence that the polymorphism was cis-acting (Flatz, 1984; Ho *et al.* 1982; Wang *et al.* 1995) (Figure 5). Relative transcript levels were used to determine the lactase persistent genotype of all 17 individuals who were heterozygous for one or more SNPs (Wang *et al.* 1995). Five of these genotyped individuals (whose DNA had been used for SNP identification) were tested for the full range of markers, and their extended haplotypes can be deduced (Figure 3a). For all other individuals, only the two most highly associated SNPs (CT-14 kb and GA-22 kb) were tested, as well as SNPs across *LCT*, which were used to deduce the likely 60 kb *LCT* haplotypes. The results were assessed in relation to lactase persistence status as well as S/L ratio. The −14 kbT allele and the −22 kbA allele were found in one or two copies in all the persistent individuals (31/36 of whom were of UK ancestry), and were not present in any of the non-persistent individuals nor in the one person for whom the diagnosis was ambiguous (Table 2) (all but one of non-UK ancestry) (Harvey *et al.* 1995b; 1998). All non-**A** haplotype chromosomes carried a C at −14 kb and a G at −22 kb, and all −14 kbT alleles and −22 kbA alleles were associated with **A** haplotype chromosomes. All 8 individuals who had been shown to be heterozygous for lactase persistence by transcript expression were heterozygous for both SNPs.

However, comparison of the CT -14 kb and GA −22 kb results with the previously determined trimodal distribution of enzyme activity ratios gave an unexpected distribution. If the T allele is truly causal of high lactase expression in adult life, it would be expected that

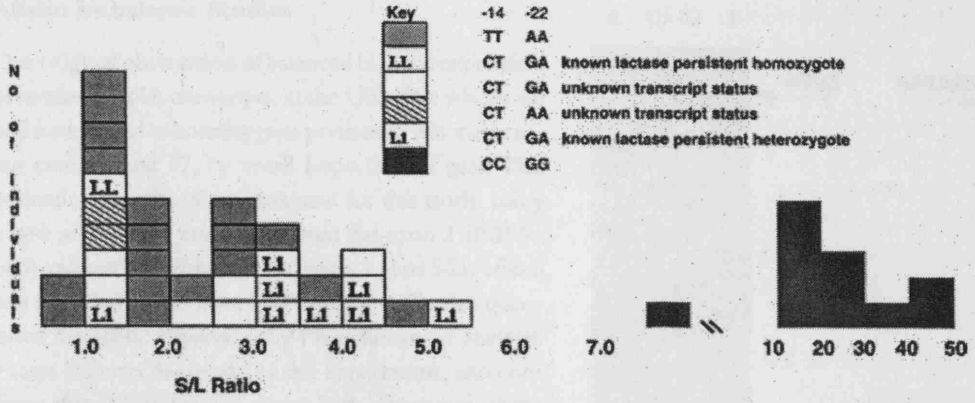


Figure 5 Histogram showing the CT –14 kb genotypes superimposed on the activity distribution, expressed as sucrase/lactase ratio (S/L) as reported before (Harvey *et al.* 1995b). Two CT samples shown cross- hatched are those for which there was discordance between the –14 kb result and the –22 kb result.

most of the homozygotes would fall in the low S/L ratio peak, while most of the CT heterozygotes would be in the intermediate peak, as had been observed previously for heterozygotes diagnosed from transcript expression studies (Wang *et al.* 1995). However, the enzyme activity ratios of the total CT heterozygote population overlapped dramatically with the TT population, and the overall difference between the two groups was not statistically significant ($p = 0.085$). This can be seen clearly in Figure 5. When individuals heterozygous for the –14 kb CT SNP were grouped according to known and unknown lactase persistence genotype status, based on RNA studies, the two groups differed from each other and it was clear that the CT 'unknown status' group did not show evidence of higher ratios ($p = 0.76$, Mann-Whitney) while the CT 'known status' group *was* significantly different from the TT homozygotes ($p = 0.006$, Mann-Whitney). The difference between the two CT groups was confirmed as being due to differences in the lactase rather than the sucrase, when the activities were compared separately (Table 2). Interestingly, the two lactase-persistent individuals who were –14 kb CT and –22 kb AA both showed the higher lactase activity (lower ratios) predicted for persistent homozygotes (Table 2, Figure 5).

The second very clear observation was that the one individual who was heterozygous **AB** haplotype and homozygous for lactase persistence, as assessed by transcript expression, was heterozygous for both CT and GA (Figure 5).

Table 2 Lactase and sucrase activities in the series of London patients, grouped according to persistence and SNP status. Monoallelic or biallelic expression was used to interpret heterozygosity or homozygosity of the phenotypic trait of lactase persistence

SNP and Persistence status	Sucrase[1] (+/−SD)	Lactase[1] (+/−SD)	S/L ratio	n
CC non-persistent	5.7 +/−2.7	0.28 +/−0.13	24.8 +/−12.3	11
CC uncertain status	13.8	1.8	7.7	1
CT persistent known heterozygote	7.9 +/−3.5	2.3 +/−0.56[2]	3.4 +/−1.1[3,4]	8
CT persistent unknown status	6.9 +/−2.4	4.1+/− 1.8[2]	2.1 +/−1.1[4]	8
TT persistent	5.9 +/−4.5	3.8 +/−3.5	1.9 +/−1.0[3]	19
CT persistent homozygote	5.2	4.8	1.1	1
CT persistent, AA at –22 kb	5.0, 4.3	5.9, 5.9	1.2, 1.4	2

[1]Activities per min. per gram wet tissue. Note that protein determinations were not done, in order to conserve sufficient biopsy material for immunohistology, electrophoresis and RNA studies, and this probably accounts for the wider scatter of the data when expressed this way. [2]$p = <.03$, Mann-Whitney and t test. [3]$p = 0.006$, [4]$p = 0.06$, Mann Whitney test. There was no significant difference in the sucrase activities between any of the groups.

Allelic Imbalance Studies

Our original observation of balanced biallelic expression of lactase mRNA transcripts, in the UK adult whom we had interpreted as homozygous persistent, was made using exons 2 and 17, by visual inspection of gels. This evaluation was therefore reassessed for this study, using a new set of PCR primers to retest the exon 2 SNP (nt 666) and testing a third SNP in exon 1 at nt 563, which was included in the same PCR product. We also quantified the allelic expression by Phosphorimage analysis. Figure 6 shows the results of this experiment, and confirms that this individual shows high expression of the **B** transcript that carries a C at –14 kb and G at –22 kb, and that the high expression of this transcript cannot therefore be due to alleles at either of these sites.

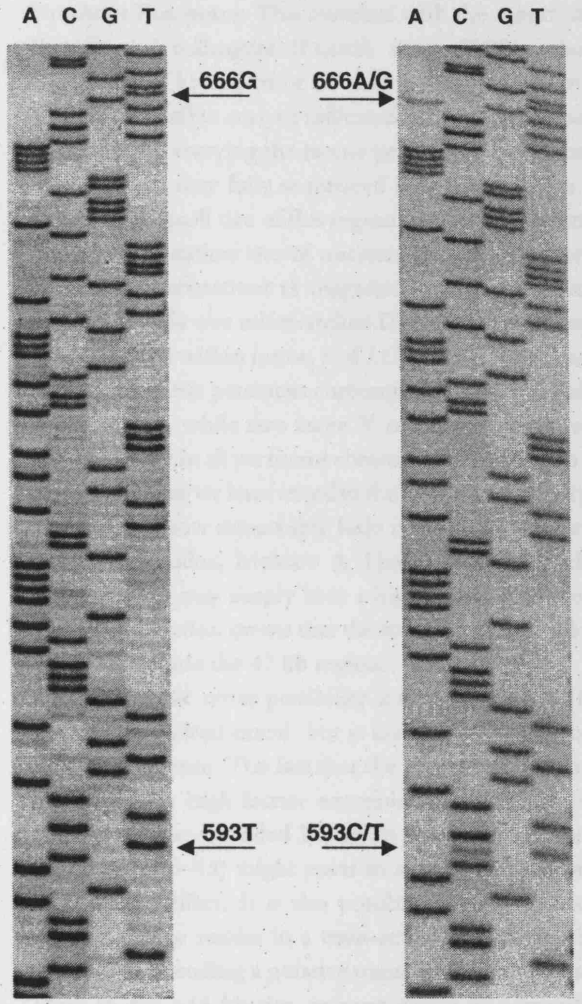


Figure 6 A ^{33}P sequencing gel covering SNPs in exon 1 and exon 2 of *LCT*, showing the relative allelic transcript expression in two individuals heterozygous at both these sites as determined from their genomic DNA. The individual on the left showed 'monoallelic' expression. The individual on the right shows high expression of both transcripts. Phosphorimage analysis showed that the relative band intensities were 56% C for exon 1 C593T, 53% A for exon 2 G666A in this individual, in contrast to 3% C for exon 1 and 0% A for exon 2 in the individual on the left. The C allele and A allele are associated with the **B** haplotype. Details of technique and controls can be found in Wang *et al.* (1998).

Discussion

In this study we have shown that *LCT* is located within a region of very high linkage disequilibrium, and many polymorphic sites show a high level of association with lactase persistence/lactose digestion. However, the association with CT-14 kb and GA-22 kb is high enough to consider them as very serious candidates for the causal mutational change. The SNPs at –14 kb and –22 kb clearly resulted from mutations on an **A** haplotype chromosome, C and G representing the ancestral alleles. The pattern of association of these two loci would suggest that the CT-14 kb SNP is due to a more recent mutation than GA-22 kb, and that they both occurred after the deletion event in intron 1. The shorter blocks of haplotype identity seen in the ancestral **A** haplotype chromosomes is consistent with the much greater relative age of the 60 kb **A** haplotype than that of the –14 kbT mutation.

Our analysis of the Finnish cohort, which shows 98% association of lactose digestion with T –14 kb, is consistent with this being causal of lactase persistence since careful examination of the different lactose digestion measures highlights the difficulties inherent in making this diagnosis. The one discrepant individual would have been differently classified (as a non-digester) if breath hydrogen alone had been used. It is however noteworthy that if the data are re-analysed using breath hydrogen alone, one other different individual becomes an exception. In contrast, the one additional exception with the GA–22 kb SNP tends to support the notion that this is *not* the causal change. This individual is unambiguously lactase non-persistent but carries an A allele, and is in agreement with the discrepant samples described by Enattah and colleagues (Enattah *et al.* 2002).

The data obtained from duodenal biopsies from individuals of non-Finnish origin also indicate that the T allele of the CT−14 kb SNP is highly associated with persistence in non-Finnish Europeans, and in all cases the data suggest that it occurs on the background of the **A** haplotype. The persistent samples we tested came mainly from Northern Europe, though 3 were from Southern Europe and 2 from the Indian sub-continent. However, the lactase activities do not correspond well with predicted genotype, except for those CT heterozygotes already known to be heterozygous for expression level who show intermediate activities. This indicates that the association of −14 kbT with expression level is not as high as in the Finnish population. If the T allele is causal of lactase persistence, it would appear that there is additional unseen heterogeneity. This could for example include the SNP at −958, which we have previously shown to affect DNA binding and which has been shown by others to alter *LCT* promoter function (Chitkara *et al.* 2001; Hollox *et al.* 1999).

Furthermore, we have previously shown that persistence is occasionally found on non-**A** haplotype chromosomes (Harvey *et al.* 1998). If presence of a T allele at the CT−14 kb SNP is causal of persistence, the most likely explanation for this would have been the occurrence of an ancestral recombination between this position and *LCT*. However, the high expressing **B** haplotype chromosome studied here in detail did not show evidence of any recombination in >420 kb upstream of *LCT* and importantly did not carry the T allele. Preliminary analysis of 36 samples from unrelated Italians described previously (Harvey *et al.* 1998) shows that while all 25 of the definite non-persistent individuals are homozygous CC, there are two other cases in which the T allele was *not* present on high expressing non-**A** haplotype chromosomes. Full analysis of the haplotype background of these samples is in progress.

In conclusion, if lactase persistence/non-persistence is due to a simple biallelic polymorphism, as originally thought, our data might suggest that *another* change is responsible for lactase persistence, and that this arose on an **A** haplotype chromosome after GA-22 kb and before CT-14 kb. Analysis of both the CEPH families and the Finnish samples shows that the three common European SNP haplotypes extend some 1000 kb in many Northern Europeans. This contrasts with the report of Enattah and colleagues (Enattah *et al.* 2002), who showed a 200 kb region of LD and whose 'identity by state' microsatellite analysis indicated a shared haplotype of only 47 kb carrying the lactase persistence allele, the region which they fully sequenced in genotyped individuals. The small size of this region may in part reflect the higher mutation rate of microsatellites, rather than ancient recombinations as suggested. For example, in their study only one microsatellite D2S3013 (see Figure 1 & 2) located within intron 1 of *LCT* breaks down the haplotypes of the persistent chromosomes at the 3′ end of the region, while two more 3′ microsatellites show the same allele in all persistent chromosomes tested. Microsatellites that we have tested in the immediate vicinity of the gene show remarkably little more diversity than the SNPs (Hollox, Mulcare & Thomas, unpublished) and D2S3013 may simply have a higher mutation rate. This interpretation means that the causal element could reside far outside the 47 kb region.

However, the other possibility is that the C-14 kbT mutation is indeed causal, but is not the only cause of lactase persistence. The fact that the **B** haplotype chromosome with high lactase expression studied here in detail shows an extended **B** haplotype over more than 700 kb (sites 3–15) might point to a separate mutation with similar effect. It is also possible that this second genetic change resides in a trans-acting element, such as the gene encoding a putative transcription factor that binds to the −14 kb site, or even that a non-genetic cause has unusually allowed adult expression of this allele. In due course these possibilities should be resolved by functional studies, but this report illustrates the difficulty of linkage disequilibrium mapping of functional nucleotide changes in regions of high LD, where the phenotype is not trivial to determine, even when there is clear evidence of a monogenic trait.

Another important conclusion from this work is that the very extended *LCT* haplotypes (particularly the **A** haplotype carrying the T allele at -14 kb), and the exceptionally long region of LD, are consistent with the notion of recent selection (Sabeti *et al.* 2002; Slatkin, 2000; Slatkin & Bertorelle, 2001) and also with our previous population genetic analysis of a smaller *LCT* haplotype, which suggested a model of recent directional selection for lactase persistence (Hollox *et al.* 2001).

Acknowledgements

The authors are grateful to Nabila Mahfiche, Janki Shah, Fella Hammachi, Lupe Polanco and Helen Bond for their contributions to this work, to the Rank Prize summer studentship funds for supporting NM, FH and HB, to Sue Povey and Mark Thomas for helpful discussion. This work was funded by the Medical Research Council through the MRC Human Biochemical Genetics Unit and the Digestive Disorders Foundation (CBH). CM is funded by a BBSRC CASE studentship.

References

Boll, W., Wagner, P. & Mantei, N. (1991) Structure of the chromosomal gene and cDNAs coding for lactase-phlorizin hydrolase in humans with adult-type hypolactasia or persistence of lactase. *Am J Hum Genet* **48**, 889–902.

Chitkara, D. K., Krasinski, S. D., Grand, R. J. & Montgomery, R. K. (2001) Regulation of human lactase phlorizin hydrolase (LPH) gene by proteins binding to sites 5′ to the Alu sequence. *Gastroenterology* **120**, A304.

Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J. M. & White, R. (1990) Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575–7.

Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L. & Jarvela, I. (2002) Identification of a variant associated with adult-type hypolactasia. *Nat Genet* **30**, 233–7.

Escalante, C. & Yang, D. C. (1993) Expression of human aspartyl-tRNA synthetase in Escherichia coli. Functional analysis of the N-terminal putative amphiphilic helix. *J Biol Chem* **268**, 6014–23.

Flatz, G. (1984) Gene-dosage effect on intestinal lactase activity demonstrated in vivo. *Am J Hum Genet* **36**, 306–10.

Harvey, C. B., Hollox, E. J., Poulter, M., Wang, Y., Rossi, M., Auricchio, S., Iqbal, T. H., Cooper, B. T., Barton, R., Sarner, M., Korpela, R. & Swallow, D. M. (1998) Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism. *Ann Hum Genet* **62**, 215–223.

Harvey, C. B., Pratt, W., Islam, I., Whitehouse, D. B. & Swallow, D. M. (1995a) DNA polymorphisms in the lactase gene: linkage disequilibrium across the 70 kb region. *Eur J Hum Genet* **3**, 27–41.

Harvey, C. B., Wang, Y., Darmoul, D., Phillips, A., Mantei, N. & Swallow, D. M. (1996) Characterisation of a human homologue of a yeast cell division cycle gene, MCM6, located adjacent to the 5′ end of the lactase gene on chromosome 2q21. *FEBS Letters* **398**, 135–140.

Harvey, C. B., Wang, Y., Hughes, L. A., Swallow, D. M., Thurrell, W. P., Sams, V. R., Barton, R., Lanzon-Miller, S. & Sarner, M. (1995b) Studies on the expression of intestinal lactase in different individuals. *Gut* **36**, 28–33.

Ho, M. W., Povey, S. & Swallow, D. (1982) Lactase polymorphism in adult British natives: estimating allele frequencies by enzyme assays in autopsy samples. *Am J Hum Genet* **34**, 650–7.

Hollox, E. J., Poulter, M., Wang, Y., Krause, A. & Swallow, D. M. (1999) Common polymorphism in a highly variable region upstream of the human lactase gene affects DNA-protein interactions. *Eur J Hum Genet* **7**, 791–800.

Hollox, E. J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A. I. & Swallow, D. M. (2001) Lactase haplotype diversity in the Old World. *Am J Hum Genet* **68**, 160–172.

Jarvela, I., Sabri Enattah, N., Kokkonen, J., Varilo, T., Savilahti, E. & Peltonen, L. (1998) Assignment of the locus for congenital lactase deficiency to 2q21, in the vicinity of but separate from the lactase-phlorizin hydrolase gene. *Am J Hum Genet* **63**, 1078–85.

Peuhkuri, K., Poussa, T. & Korpela, R. (1998) Comparison of a portable breath hydrogen analyser (Micro H2) with a Quintron MicroLyzer in measuring lactose maldigestion, and the evaluation of a Micro H2 for diagnosing hypolactasia. *Scand J Clin Lab Invest* **58**, 217–24.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., D., Cooper, R., Kwiatkowski, D., Ward, R. & Lander, E. S. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–7.

Slatkin, M. (2000) Balancing selection at closely linked, overdominant loci in a finite population. *Genetics* **154**, 1367–78.

Slatkin, M. & Bertorelle, G. (2001) The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics* **158**, 865–74.

Swallow, D. M. & Hollox, E. J. (2000) The genetic polymorphism of intestinal lactase activity in adult humans. In: Scriver, C. R., Beaudet, A. L., Sly, W. S., Valle, D. (eds) The Metabolic and Molecular Bases of Inherited Disease. McGraw-Hill, New York.

Thomas, M. G., Bradman, N. & Flinn, H. M. (1999) High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. *Hum Genet* **105**, 577–81.

Wang, Y., Harvey, C., Rousset, M. & Swallow, D. (1994) Expression of intestinal mRNA transcripts during development: analysis by a semi-quantitative RNA PCR method. *Ped Res* **36**, 514–521.

Wang, Y., Harvey, C. B., Hollox, E. J., Phillips, A. D., Poulter, M., Clay, P., Walker-Smith, J. A. & Swallow, D. M. (1998)

The genetically programmed down-regulation of lactase in children. *Gastroenterology* **114**, 1230–1236.

Wang, Y., Harvey, C. B., Pratt, W. S., Sams, V. R., Sarner, M., Rossi, M., Auricchio, S. & Swallow, D. M. (1995) The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element. *Hum Mol Genet* **4**, 657–662.

Wegner, S. A., Ehrenberg, P. K., Chang, G., Dayhoff, D. E., Sleeker, A. L. & Michael, N. L. (1998) Genomic organization and functional characterization of the chemokine receptor CXCR4, a major entry co-receptor for human immunodeficiency virus type 1. *J Biol Chem* **273**, 4754–60.

Ye, S., Dhillon, S., Ke, X. & Collins A. R., Day IN (2001) An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res* **29**, E88–8.

Received: 24 October 2002
Accepted: 31 January 2003

Am. J. Hum. Genet. 74:1102–1110, 2004

The T Allele of a Single-Nucleotide Polymorphism 13.9 kb Upstream of the Lactase Gene (*LCT*) (*C−13.9kbT*) Does Not Predict or Cause the Lactase-Persistence Phenotype in Africans

Charlotte A. Mulcare,[1,2] Michael E. Weale,[1,*] Abigail L. Jones,[1] Bruce Connell,[3] David Zeitlyn,[4] Ayele Tarekegn,[1] Dallas M. Swallow,[2] Neil Bradman,[1] and Mark G. Thomas[1]

[1]The Centre for Genetic Anthropology (TCGA) and [2]Galton Laboratory, University College London, London; [3]Department of Languages, Literatures, and Linguistics, York University, Toronto; and [4]Centre for Social Anthropology and Computing, Department of Anthropology, University of Kent, Canterbury, United Kingdom

The ability to digest the milk sugar lactose as an adult (lactase persistence) is a variable genetic trait in human populations. The lactase-persistence phenotype is found at low frequencies in the majority of populations in sub-Saharan Africa that have been tested, but, in some populations, particularly pastoral groups, it is significantly more frequent. Recently, a CT polymorphism located 13.9 kb upstream of exon 1 of the lactase gene (*LCT*) was shown in a Finnish population to be closely associated with the lactase-persistence phenotype (Enattah et al. 2002). We typed this polymorphism in 1,671 individuals from 20 distinct cultural groups in seven African countries. It was possible to match seven of the groups tested with groups from the literature for whom phenotypic information is available. In five of these groups, the published frequencies of lactase persistence are ≥25%. We found the T allele to be so rare that it cannot explain the frequency of the lactase-persistence phenotype throughout Africa. By use of a statistical procedure to take phenotyping and sampling errors into account, the T-allele frequency was shown to be significantly different from that predicted in five of the African groups. Only the Fulbe and Hausa from Cameroon possessed the T allele at a level consistent with phenotypic observations (as well as an Irish sample used for comparison). We conclude that the *C−13.9kbT* polymorphism is not a predictor of lactase persistence in sub-Saharan Africans. We also present Y-chromosome data that are consistent with previously reported evidence for a back-migration event into Cameroon, and we comment on the implications for the introgression of the *−13.9kb*T* allele.

Introduction

Lactose, a disaccharide, is the principal calorific component of milk. To be absorbed, it must be hydrolyzed, a reaction mediated by the enzyme lactase (lactase-phlorizin hydrolase). In most mammals that have been studied, the level of the lactase enzyme is severely reduced some time after weaning, so adults cannot digest lactose effectively (reviewed in Swallow and Hollox 2000). However, in humans, the ability to digest lactose throughout adulthood (lactase persistence [MIM 223100]) exists as a dominant Mendelian polymorphic trait (Sahi 1974; Swallow and Harvey 1993).

Received December 2, 2003; accepted for publication March 10, 2004; electronically published April 20, 2004.

Address for correspondence and reprints: Dr. Dallas Swallow, Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE. E-mail: dswallow@hgmp.mrc.ac.uk.

* This author devised the statistical analysis method used in this study; the program is available on the TCGA Web site (see the "Electronic-Database Information" section).

0002-9297/2004/7406-0004$15.00

Lactase persistence varies widely in frequency among different human populations, both between and within continents. It is generally found at high frequencies in populations of European descent, in which, for example, Dutch and Swedish studies recorded frequencies of 100% and 99%, respectively (reviewed in Swallow and Hollox 2000). Lactase-nonpersistent individuals (lactose nondigesters) may suffer adverse symptoms from milk ingestion resulting from the breakdown of lactose by bacteria in the gut, varying from mild flatulence to severe abdominal pains and diarrhea.

Although the structure and full exonic sequence of the lactase gene (*LCT* [MIM 603202]) has been known since 1991 (Boll et al. 1991), the causative mechanism for lactase persistence has proved more elusive. Recently, Enattah and colleagues (2002) showed that the T allele of a C/T transition 13.9 kb upstream from exon 1 of *LCT* in intron 13 of the gene *MCM6*, here referred to as *−13.9kb*T*, was completely associated with lactase persistence in a sample of 196 unrelated Finnish individuals, whose diagnoses were made from intestinal biopsy specimens (lactase persistence: $n = 137$; lactase nonpersistence: $n = 59$). Furthermore, 40 lactase-non-

persistent individuals from various populations (Germany, Italy, South Korea) were all homozygous for the C allele. Finally, DNA from individuals of unknown phenotype collected in Finland (n = 938), France (n = 17), and the United States (two data sets: African descent [n = 96] and European descent [n = 92]) had frequencies of the CC genotype that appeared to be consistent with the frequencies of lactase nonpersistence reported for those groups. Although no functional mechanism was shown at that time, presence of *−13.9kb*T* was proposed as a robust marker for lactase persistence. Very recent studies have suggested that this SNP is located in an enhancer element and that the two alleles show some difference in function (Olds and Sibley 2003; Troelsen et al. 2003). *C−13.9kbT* typing is now being offered as a genetic test for lactase persistence in Finland (Medix Laboratory), where the strong correlation between the T allele and lactase persistence was first reported, and such testing is being considered for use elsewhere (Buning et al. 2003; Hoegenauer et al. 2003).

To date, there have been no reports of allele frequencies for the *C−13.9kbT* polymorphism in populations living in Africa. Although the *−13.9kb*T* allele frequency in Americans with African ancestry is consistent with their lactase-persistence frequency (Enattah et al. 2002), there is known to be substantial admixture between African Americans and European Americans (Parra et al. 1998). Previous studies of African populations showed variation in the frequency of lactase persistence among population groups, as well as a complex pattern of distribution (reviewed in Flatz 1987; Holden and Mace 1997; Swallow and Hollox 2000). Pastoralists, such as the Fulbe in Nigeria, typically have higher frequencies of lactase persistence than nonpastoralists in the same country—for example, the Yoruba and Igbo (Kretchmer et al. 1971; Olatunbosun and Adadevoh 1971; Ransome-Kuti et al. 1972, 1975; Flatz 1987; Holden and Mace 1997). The lactase-persistence phenotype is usually observed at low frequencies in Bantu- and Khoisan-speaking groups (<20%) (Cook and Kajubi 1966; Cook et al. 1967, 1973; Cox and Elliott 1974; Nurse and Jenkins 1974; O'Keefe and Adam 1983; Segal et al. 1983; O'Keefe et al. 1984).

We have typed *C−13.9kbT* in 1,671 individuals from 20 different African populations, including both milk-drinking and non–milk-drinking groups. Since phenotype data was not available for these samples, we performed an ethnologically matched group study to determine whether *−13.9kb*T* was associated with lactase persistence in seven African samples and in one northern European sample (to provide a comparative group), using a statistical procedure to take both sampling and phenotyping error into account. We show that, in most cases in Africa, the frequency of *−13.9kb*T* is too low to explain the observed frequency of lactase persistence.

Material and Methods

Samples

DNA was extracted from buccal swabs collected from males belonging to different groups living in various regions of Africa, including populations that have a history of pastoralism and milk drinking and others that do not (table 1). DNA was also obtained from a sample of unrelated Irish individuals for comparative purposes. Informed consent was obtained from all donors. Ethical approval was obtained from University College Hospitals and University College London Joint Committee on the Ethics of Human Research (reference number 99/0196). Appropriate permissions were obtained in each of the collection countries (reference number for Cameroon 0093MINREST/B00/D00/D10/D12). Each donor provided biographical details, such as self-defined ethnic identity, first and second language, and place of birth, with similar information on his mother, father, maternal grandmother, and paternal grandfather. Individuals were classified as belonging to a given cultural group if their own self-declared identity concurred with that they ascribed to both mother and father. Where there were <10 individuals of the same declared cultural identity, they were classified as "other." Individuals with partially unknown or mixed ancestry at the parental level were also classified as "other." The Ethnographic Atlas (Murdock 1967) and a summary table of pastoralists (Blench 1999) were used to obtain information about pastoralism and milk practice, and the Ethnologue Web site was used to check possible linguistic relationships between groups.

Selection of Matched Populations for Which Phenotypic Data Were Available

For many of the population groups, a matching population sample with lactose-tolerance (digestion) data could be found in the literature. Matching samples were selected to fulfill the following criteria: (1) same declared cultural identity and (2) residency in the same country or in a neighboring country. Sources for data on lactase persistence are listed in table 2.

Typing the C−13.9kbT *Polymorphism*

PCR primers LAC-C-M-U (5′-GCTGGCAATACAGATAAGATAATGGA-3′) and LAC-C-L2 (5′-CTGCTTTGGTTGAAGCGAAGAT-3′) were designed to amplify the region containing the C/T polymorphism (Enattah et al. 2002). The penultimate base of the LAC-C-M-U primer (G) introduces a base change such that the PCR product will be cut by *Hin*fI when the T allele is present, giving digestion product sizes of 177 bp and 24 bp, but

not when the C allele is present, giving a digestion product size of 201 bp. PCR reactions were performed in a total volume of 10 μl containing 200 μM dNTPs, 10 mM Tris-HCl (pH 9.0), 0.1% Triton X-100, 0.01% gelatin, 50 mM KCl, 2.0 mM $MgCl_2$, 0.13 U *Taq* polymerase enzyme (HT Biotech), 9.3 nM *Taq*Start monoclonal antibody (BD Biosciences, Clontech), and 0.3 μM primers. The *Taq* and *Taq*Start monoclonal antibodies were premixed prior to being added to the other reagents. Thermal cycling conditions were an initial denaturation stage at 95°C for 5 min, then 35 cycles of 95°C for 1 min, 59°C for 1 min, and 72°C for 1 min, followed by a final elongation stage at 72°C for 5 min. Digestions were performed at 37°C overnight in the original PCR plate in a total volume of 25 μl. Each reaction contained the entire PCR product, 0.25 U of *Hin*fI, 0.01 μg/μl acetylated BSA, and New England Biolabs Buffer 2, as recommended by the manufacturer. The digestion products were run on a 3% agarose gel, and DNA bands were visualized by use of ethidium bromide staining. Gel phenotypes showing a single band of 201 bp (C) or 177 bp (T) were interpreted as genotypes $-13.9kb*CC$ and $-13.9kb*TT$, respectively.

Genotype Error Checking

One positive control (a known CT heterozygote) and a blank were included in every 96-well plate. In addition, a set of 50 randomly selected samples were retyped "blind." All controls and the 50 retyped samples matched the initial typing. As a further control that the results for the SNP matched those reported by Enattah et al. (2002), our SNP protocol was used for typing phenotyped Finnish individuals. There was a high, although incomplete, correlation between the T allele and lactose-tolerance test results (Poulter et al. 2003). The level of discrepancy was attributable to inaccuracies of the tolerance testing and was consistent with the error rates we use in our statistical procedure described below.

Statistical Analysis

Our genotype-error–checking procedure (see above) suggested that the genotyping error rate was zero or negligible. However, phenotyping error in the determination of lactose digestion as an indirect test for lactase-persistence status (e.g., by measurement of breath hydrogen or blood glucose) is known to occur at appreciable levels. It is the convention in the medical literature to describe lactose digesters as "negative" and nondigesters as "positive"—that is, giving a positive diagnosis in a lactose-tolerance test. We obtained information on the error rates of false-negative (FN) results (i.e., when a nonpersistent person appears to be a digester) and false-positive (FP) results (i.e., when a persistent person appears to be a nondigester) from three studies in which the correct lactase phenotype was ascertained by peroral jejunal biopsy and from two studies in which the "correct" phenotype was determined by the "gold standard" method. When the "gold standard" method is used, at least two of three separate noninvasive tests (namely, blood glucose, breath hydrogen, and urine galactose) must concur. Error rates for the three studies ascertained by biopsy were as follows (sample sizes in the denominator): (1) blood glucose FN = 6/25 and FP = 1/25, breath hydrogen FN = 0/25 and FP = 0/25 (Newcomer et al. 1975); (2) blood glucose FN = 0/7 and FP = 0/8, breath hydrogen FN = 0/7 and FP = 0/8 (Howell et al. 1981); and (3) blood glucose not measured, breath hydrogen FN = 5/16 and FP = 2/47 (Arola et al. 1988). Error rates for the two studies ascertained by the "gold standard" method were as follows (sample sizes in the denominator): (1) blood glucose FN = 3/35 and FP = 3/35, breath hydrogen FN = 2/35 and FP = 2/35 (Puehkuri 2000); and (2) blood glucose FN = 1/49 and FP = 1/5, breath hydrogen FN = 2/49 and FP = 1/5 (Kurt et al. 2003). Exact protocols for the blood glucose and breath hydrogen tests varied among the five studies, as they did among the studies on matching populations reported later. However, all protocols involved the measurement of changes in plasma glucose or exhaled hydrogen at one or more time intervals between 30 min and 4 h after administration of at least 50 g lactose. We combined the results from the above five studies to perform a rough averaging over the differences in protocols used (blood glucose FN = 10/116 and FP = 5/73, breath hydrogen FN = 9/132 and FP = 5/120). The combined results suggest that, even if a population has no lactase-persistent individuals, we would expect, by use of one of these two methods, to find between 5% and 10% FNs (i.e., apparent lactose digesters).

Given the above and assuming that the underlying phenotyping error rates acting in these independent studies are applicable to other studies that use the same measurement techniques, we devised a statistical procedure that allowed us to test whether the frequency of lactose digesters predicted by the $C-13.9kbT$ genotype data was sufficient to explain the observed frequency found in the phenotyped group. We took both phenotyping error and four possible sources of sampling uncertainty into account: (1) sampling uncertainty in p, the frequency of the $-13.9kb*T$ in the genotyped group; (2) sampling uncertainty in f_n, the frequency of false negatives according to the phenotyping method used; (3) sampling uncertainty in f_p, the frequency of false positives according to the phenotyping method used; and (4) sampling uncertainty in L_{app}, the frequency of apparent lactase persistence in the phenotyped group.

The procedure was performed as follows:

1. A value for p was drawn from a Beta(T+1, C+1)

Table 1

Published Information on Pastoralism and the Observed Frequency of *−13.9kb*T*, by Population Group

Country and Group	Nomadic Pastoralist[a] Status	% Dependent on Animal Husbandry (Milking Status)[b]	No. of Individuals	No. of Individuals with Genotype			Observed Frequency of *−13.9kb*T*
				CC	CT	TT	
Cameroon:							
Fulbe[c]	Yes	46–55% (Yes)	49	39	9	1	.112
Hausa[d]	No	16–35% (No)	18	14	3	1	.139
Kwanja[e]	No	NA	70	70	0	0	0
Mambila[e]	No	16–25% (No)	122	121	1	0	.004
Nso[e] (Nsaw)	No	6–15% (No)	126	126	0	0	0
Yamba[e]	No	NA	21	21	0	0	0
Other	NA	NA	128	118	9	1	.043
Nigeria:							
Ibibio	No	6–15% (No)	110	110	0	0	0
Oron	No	6–15% (No)[i]	44	44	0	0	0
Other	No	NA	22	22	0	0	0
Malawi:							
Chewa[e]	No	6–15% (Yes)	84	84	0	0	0
Ngoni[e]	No	6–15% (Yes)	14	14	0	0	0
Tumbuka[e]	No	6–15% (No)	58	58	0	0	0
Yao[e]	No	6–15%	49	49	0	0	0
Other[e]	No	NA	58	58	0	0	0
Senegal:							
Wolof[f]	No	26–35% (Yes)	69	69	0	0	0
Manjak	No	NA	93	93	0	0	0
Other	NA	NA	19	18	1	0	.026
Sudan (North):							
Ga'ali	No	NA	30	30	0	0	0
Shaigi	No	NA	11	11	0	0	0
Other[g]	Mixed	NA	88	88	0	0	0
Sudan (South):							
Dinka	Yes	46–55% (Yes)	34	34	0	0	0
Nuer	Yes	46–55% (Yes)	13	13	0	0	0
Other[g]	Mixed	NA	73	73	0	0	0
Ethiopia:							
Nuer	Yes	46–55% (Yes)	119	119	0	0	0
Anuak (Anywak)	Yes[h]	6–15% (Yes)	108	108	0	0	0
Other	NA	NA	1	1	0	0	0
Uganda:							
Mussesse[e]	No	NA	22	22	0	0	0
Other[e]	No	NA	18	18	0	0	0
Ireland	NA	36–45% (Yes)	47	1	10	36	.872

Note.—NA = not available.

[a] Pastoralists who migrate with their animals. From table 2.1 in Blench (1999).

[b] Murdock (1967).

[c] Fulbe with a sedentary lifestyle.

[d] Immigrant population from Nigeria.

[e] Bantoid-language speakers.

[f] Includes 23 "Lebou" individuals. Lebou is a dialect of Wolof.

[g] This group includes 12 individuals from traditional milk-drinking peoples of known high frequency for lactase persistence (Beja, Misseri, Gomoeia, and Shilluk [Bayoumi et al. 1981, 1982]).

[h] Probably do not drink fresh milk (A. Tarekegn, unpublished data).

[i] Identical to Ibibio in this respect.

distribution, where T is the number of T alleles and C is the number of C alleles found in the genotyped group. This beta distribution describes the posterior distribution for p, given the genotype data, assuming a Uniform(0,1) prior.

2. The predicted frequency of true lactase persistence in the population, L_{true}, was calculated as $p^2+2p(1-p)$ (i.e., the expected frequency of TT + CT genotypes under Hardy-Weinberg equilibrium).
3. Values for f_n and f_p were drawn from Beta(11,107)

Table 2

Comparisons with Published Lactose-Digester Frequencies in Matching Populations, Taking into Account Sampling and Phenotyping Error

Group	Country of Genotyped Sample (No.)	Expected Frequency of Lactose Digesters	Country of Phenotyped Sample (No.)	Test Method	Observed Frequency of Lactose Digesters in Phenotyped Sample	Reference	*P* Value
Fulbe[a]	Cameroon (n = 49)	.265	Nigeria (n = 24)	Blood glucose	.292	Kretchmer et al. 1971	1
Hausa	Cameroon (n = 18)	.305	Nigeria (n = 17)	Blood glucose	.235	Kretchmer et al. 1971	.749
Wolof	Senegal (n = 69)	.086	Senegal (n = 53)	Blood glucose	.509	Arnold et al. 1980	0
Ga'ali (Jaali)	Sudan (North) (n = 30)	.068	Sudan (n = 113)	Breath hydrogen	.531	Bayoumi et al. 1981	0
Shaigi	Sudan (North) (n = 11)	.068	Sudan (n = 42)	Breath hydrogen	.381	Bayoumi et al. 1981	.025
Nuer	Sudan (South), Ethiopia (n = 132)	.068	Sudan (n = 23)	Breath hydrogen	.217	Bayoumi et al. 1982	.030
Dinka	Sudan (South) (n = 34)	.068	Sudan (n = 208)	Breath hydrogen	.255	Bayoumi et al. 1982	.001
European	Ireland (n = 47)	.918	Ireland (n = 50)	Blood glucose	.900	Fielding et al. 1981[b]	1

NOTE.—Expected frequency of lactose digesters, taking into account the test error rate by the method used in the matched population = $L_{true}(1-f_p) + (1-L_{true})f_n$, where L_{true} = frequency of CT + TT genotypes assuming Hardy-Weinberg equilibrium, (f_n, f_p) = (10/116, 5/73) if the blood glucose test method is used and (f_n, f_p) = (9/132, 5/120) if the breath hydrogen test method is used. *P* value = result of test described in the "Statistical Analysis" section.

[a] Fulbe with a sedentary lifestyle.

[b] Blood glucose results only taken from this source, by use of a rise of >20 mg/dl to define lactose digester.

and Beta(6,69) distributions, respectively, if phenotyping was by the blood glucose method and from Beta(10,124) and Beta(6,116) distributions, respectively, if phenotyping was by the breath hydrogen method. Again, these beta distributions describe the posterior distribution for f_n and f_p, given the combined false error rate data reported above and assuming a Uniform(0,1) prior.

4. The predicted frequency of apparent lactose digesters accounting for phenotyping error, L_{app}, was calculated as $L_{true}(1-f_p) + (1-L_{true})f_n$.
5. A simulated value for n_L, the number of lactose digesters observed in the phenotyped group was drawn from a Binomial(n,L_{app}) distribution, where n is the number sampled in the phenotyped group.
6. Steps 1–5 were repeated 100,000 times (N = 100,000) to build up a Monte Carlo sampling distribution for n_L under the null hypothesis that the C/T genotype and phenotyping error alone account for the apparent frequency of lactose digesters.
7. Let S_g be the sum of simulated n_L values greater than or equal to the observed n_L value, and let S_l be the sum of simulated n_L values less than or equal to the observed n_L value. A two-tailed P value for the observed n_L under the null hypothesis was found as $2 \times \min(S_g,S_l)/N$.

Y-Chromosome Haplotypes

Cruciani and colleagues (2002) reported the presence of a non-African Y-chromosome lineage (M173-derived haplotype 117, or R1*, by use of nomenclature of the Y-Chromosome Consortium [2002]) in population groups from northern Cameroon. These authors suggested this was due to a back-migration event from outside sub-Saharan Africa. We typed our samples for a genealogically similar marker, 92R7 (Mathias et al. 1994), which is ancestral to M173 but for which intermediate haplotypes (92R7-derived, M173-ancestral) have not been reported in any study of sub-Saharan African populations to date. We also typed 92R7-derived samples for six Y-chromosome microsatellites (DYS19, DYS388, DYS390, DYS391, DYS392, and DYS393) to investigate the intrahaplogroup diversity. Protocols for the typing of 92R7 and the microsatellites are those described by Thomas et al. (1999). Microsatellite repeat numbers were assigned according to the nomenclature of Kayser and colleagues (1997). Genetic diversity, h, and its SE were calculated according to the unbiased formulae in the work of Nei (1987). Since Y-chromosome typing was not successful in 72 of the 1,671 samples, we also report separately the different Y-chromosome sample sizes.

Results

C−13.9kbT *in Africa*

The frequency of *−13.9kb*T* was low or zero in most of the African groups tested, whereas it was very high in the Irish sample (table 1). In the African populations, the *−13.9kb*T* allele was only found in a few individuals; all but one of these individuals were from Cameroon and lived close to the same market town, Mayo Darle. Of these individuals, there were 10 Fulbe, 4 Hausa, 1 Mambila, and 10 "others" (mixed ancestry or from other ethnic groups). In all but two cases (one Hausa and one "other" individual), the *−13.9kb*T*-carrying individuals or one or both of their parents spoke Fulfulde, a Fulbe language. This association between possession of the *−13.9kb*T* allele and speaking Fulfulde was significant both in the Cameroonian sample as a whole ($P < .001$, $n = 534$, Fisher's exact test) and in the non-Fulbe Cameroonians ($P = .015$, $n = 485$, Fisher's exact test). The one individual from Senegal carrying *−13.9kb*T* allele was of mixed Wolof and Toucouleur (Tukulor) ancestry. There were no significant departures from Hardy-Weinberg equilibrium in any of the ascribed ethnic groups where T alleles were observed (by use of the method of Guo and Thompson [1992]).

It is noteworthy that *−13.9kb*T* was not found in East Africa at all, even though the data sets included many known pastoralists and groups with a high frequency of lactase persistence (table 1).

Matched Populations for Which Phenotypic Data Were Available

In some cases, it was possible to find closely matching populations in the literature that had phenotype information (table 2). Comparisons of the predicted frequencies of lactase persistence, deduced from the frequency of *−13.9kb*TT* and *−13.9kb*CT* genotypes, with the reported frequencies obtained from lactose-tolerance testing, showed these were significantly different in all of the African populations except the Fulbe and the Hausa. Only in these two Cameroonian groups was *−13.9kb*T* found at frequencies sufficient to explain the raised incidence of lactase persistence. In contrast to the generally poor correspondence between genotypic and phenotypic data in African populations, our Irish sample shows excellent correspondence between predicted and observed frequencies of lactase persistence in the genotyped and phenotyped groups, consistent with the findings of Enattah et al. (2002).

Y-Chromosome Data

The 92R7-derived haplogroup was extremely rare in the sub-Saharan African populations sampled. Most

92R7-derived chromosomes were found in Cameroon, with 8/42 in the Fulbe, 1/110 in Mambila, 1/65 Kwanja, and 5/113 "others." Outside Cameroon, we found five 92R7-derived chromosomes in northern Sudan (3/11 Shaigi, 2/29 Ga'ali) and one in southern Sudan (1/72 "others"). The microsatellite haplotype diversity of 92R7-derived chromosomes in Cameroon was high, with 10 haplotypes observed among 15 individuals (h = 0.933, SE = 0.0449, average repeat size variance = 0.224).

Discussion

The absence of *−13.9kb*T* in most of the African populations typed, which included several milk-drinking groups (table 1), suggests that it is not a reliable predictor of the lactase-persistence phenotype in populations from this region. This, in turn, indicates either that it is not a causative mutation or that it is not the sole causative mutation in all human populations.

This conclusion is consistent with a previous study (Poulter et al. 2003). In a series of 48 London patients of various ancestry, from whom intestinal biopsies were obtained, the correlations of lactase activity and sucrase/lactase ratio with *−13.9kb*CT* and *−13.9kb*TT* genotype were not as tight as might have been expected for a *cis*-acting causal change. In contrast to this, in a recent Finnish study, the *13.9kb*CT* heterozygotes did have activity intermediate between the *13.9kb*CC* and *13.9kb*TT* homozygotes (Kuokkanen et al. 2003).

Previous studies have shown that, outside Africa, there are very few common *LCT* gene haplotypes (A, B, C, and U) (Hollox et al. 2001), and recent studies in Europeans have shown that linkage disequilibrium extends over at least 1 Mb (Poulter et al. 2003), as demonstrated clearly by Bersaglieri and colleagues in this issue of the *Journal* (Bersaglieri et al. 2004 [in this issue]). The *−13.9kb*T* allele is carried on the background of the extended A haplotype that is most common in northern Europeans and may have reached high frequencies as a result of selection (Poulter et al. 2003). However, many A haplotype chromosomes do not carry T at −13.9 kb. A comparison of the occurrence of this allele, as well as alleles at other recently described loci that subdivide the A haplotype (such as *G−22kbA*, [Enattah et al. 2002]), suggests that *−13.9kb*T* is the most recent (Poulter et al. 2003). It is possible that the *C−13.9kbT* transition occurred more recently than another (as yet unknown) mutation that is the true causal change both in Africa and Europe. Recent transfection studies do, however, suggest a functional role for *C−13.9kbT* (Olds and Sibley 2003; Troelsen et al. 2003).

In a few rare individuals, high expression of the mRNA transcript, encoded by the *LCT* allele of a non-A haplotype chromosome, has been observed (Poulter et al. 2003). In particular, a single individual in a United Kingdom cohort was interpreted as being heterozygous for the A and B haplotypes, as well as for *C−13.9kbT*, and showed high expression of lactase from both transcripts, suggesting that there may be heterogeneity of the cause of lactase persistence in Europe (Poulter et al. 2003).

It seems probable that the C-to-T transition at −13.9 kb occurred in a non–sub-Saharan African population that contributed to the current population of Europe. If this were the case, then its presence in Cameroon, and especially in people of Fulbe cultural identity or with Fulfulde-speaking ancestry, could be explained by introgression from outside sub-Saharan Africa.

Our Y-chromosome data corroborate the results of Cruciani and colleagues (2002) in finding high frequencies in our Cameroonian samples of a haplogroup that is generally absent from sub-Saharan Africa. Phylogeographic arguments suggest that this haplogroup (R1*, by use of the nomenclature of the Y-Chromosome Consortium [2002]) has a non-African origin. Cruciani and colleagues (2002) found R1* Y chromosomes at an average frequency of 40% in several northern Cameroonian groups, including one Fulbe group. We found evidence for the same haplotype (typed by use of a marker that appears phylogenetically identical in this part of Africa) in our samples from central Cameroon, with a particularly high frequency (19%) in the Fulbe group that was tested. The Y-chromosome microsatellite diversity we observed indicates that this haplogroup could not have been brought to this part of Africa by a single recent founder. The origins of the Fulbe are the subject of debate, but the group is thought to be from outside Cameroon; on the basis of ethnic traditions and linguistic similarities between Fulbe languages and Tukulor (Toucouleur), an origin in the Futa Toro region of the Senegal river basin has been proposed (Newman 1995). It is possible that the back-migration event that led to the introduction of R1* into sub-Saharan Africa (Cruciani et al. 2002) also brought the *−13.9kb*T* allele and that the Fulbe of central Cameroon migrated locally from the north. However, haplogroup R1* is also found at high frequencies in several non-Fulbe groups in the Extreme North Province of Cameroon, where the *−13.9kb*T* allele is found at low frequencies (<3%, data not shown). Thus, the demographic processes leading to the presence of the *−13.9kb*T* allele in Cameroon may be not be the same as those leading to the Y-chromosome introgression but could instead relate more specifically to Fulbe migration history. Further studies on the distribution of the *−13.9kb*T* allele and of other genetic markers in this part of Africa are required to resolve this question.

We have shown that *−13.9kb*T* is not associated

with lactase persistence in a wide range of African populations. It would now be appropriate to undertake more extensive genotyping and phenotype characterization on the same individuals in multiple African groups. It will be of interest to determine whether lactase persistence is associated with the same or a different haplotype and whether such haplotypes have an extended length consistent with a recent selective sweep, as was found for Europeans (Bersaglieri et al. 2004 [in this issue]). African populations display multiple lifestyles, with milk-drinking and non–milk-drinking groups often living in close proximity, and have complex demographic histories. Understanding the genetic determinants of lactase persistence in African populations will help explain the genetic history of the lactase-persistence phenomenon and should ultimately have positive implications for public health.

Conclusion

The T allele located 13.9 kb upstream of *LCT* has been claimed by Enattah and colleagues (2002) as a predictor of lactase persistence in European populations. It does not fulfill that function in sub-Saharan Africans. Use of the *C−13.9kbT* polymorphism as a diagnostic predictor of adult hypolactasia outside Europe should therefore be approached with caution. Our results show that the *−13.9kb*T* allele cannot be causal of lactase persistence in most Africans, although it could possibly explain lactase persistence in some Cameroonians. Data presented in this study support the possibility that the presence of the *−13.9kb*T* allele in Cameroon is due to introgression from outside sub-Saharan Africa. The combined results from *C−13.9kbT* and the Y-chromosome analysis suggest a complex demographic history for this part of Africa, which includes at least one major introduction of genes from outside the region.

Acknowledgments

We thank Dominic Gormis, Esther William, Tanelli Helenius, Jim Wilson, Pieta Nasanen, John Greenhalgh, Jane Moore, Richard Phillips, Katya Bulgina, Alex Murray, Ali Barwhani, Corine Atton, and Noreen von Cramon-Taubadel, who collected and extracted many of the DNA samples used in the present study and/or tested for Y-chromosome markers. We also thank Dr. Roger Blench and Dr. Clare Holden for helpful discussions. C.A.M. was funded by a BBSRC CASE studentship.

Electronic-Database Information

The URLs for data presented herein are as follows:

Ethnologue, Languages of the World, http://www.ethnologue.com/

Medix Laboratory, http://www.medix.fi/tiedotteet/tuoteinfo_02/03.htm (for SNP testing for lactase-persistence diagnosis)

Online Mendelian Inheritance in Man (OMIM): http://www.ncbi.nlm.nih.gov/Omim/ (for lactase persistence and *LCT*)

The Centre for Genetic Anthropology (TCGA) Software Page, http://www.ucl.ac.uk/tcga/software/ (for the statistical analysis program used for the ethnologically matched group study)

References

Arnold J, Diop M, Kodjovi M, Rozier J (1980) L'intolerance au lactose chez l'adulte au Senegal (Lactose intolerance in adults in Senegal). C R Seances Soc Biol Fil 174:983–992 (In French)

Arola H, Koivula T, Jokela H, Jauhiainen M, Keyrilainen O, Ahola T, Uusitalo A, Isokoski M (1988) Comparison of indirect diagnostic methods for hypolactasia. Scand J Gastroenterol 23:351–357

Bayoumi RAL, Flatz SD, Kuhau W, Flatz G (1982) Beja and Nilotes: nomadic pastoralist groups with opposite distributions of the adult lactase phenotypes. Am J Phys Anthropol 58:173–178

Bayoumi RAL, Saha N, Salih AS, Bakkar AE, Flatz G (1981) Distribution of the lactase phenotypes in the population of the Democratic Republic of the Sudan. Hum Genet 57:279–281

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111–1120 (in this issue)

Blench R (1999) Why are there so many pastoral groups in eastern Africa? In: Azarya V, Breedveld A, De Bruijn M, Van Dijk H (eds) Pastoralists under pressure? Fulbe societies confronting change in west Africa. Brill Press, Boston

Boll W, Wagner P, Mantei N (1991) Structure of the chromosomal gene and cDNAs coding for lactase-phlorizin hydrolase in humans with adult-type hypolactasia or persistence of lactase. Am J Hum Genet 48:889–902

Buning C, Jurga J, Fiedler T, Kupferling S, Worm M, Weltrich R, Genschel J, Lochs H, Schmidt H, Ockenga J (2003) Genetic background of lactose intolerance and implications for diagnosis. Gastroenterology Suppl 124:A144

Cook G, Asp N, Dahlqvist A (1973) Lactose absorption kinetics in Zambian African subjects. Br J Nutr 30:519–527

Cook G, Kajubi S (1966) Tribal incidence of lactase deficiency in Uganda. Lancet 1:725–729

Cook G, Lakin A, Whitehead R (1967) Absorption of lactose and its digestion products in the normal and malnourished Ugandan. Gut 8:622–627

Cox J, Elliott F (1974) Primary adult lactose intolerance in the Kivu lake area: Rwanda and the Bushi. Am J Dig Dis 19:714–724

Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet 70:1197–1214

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I (2002) Identification of a variant associated with adult-type hypolactasia. Nat Genet 30:233–237

Fielding J, Harrington M, Fottrell P (1981) The incidence of primary hypolactasia amongst the Irish. Ir J Med Sci 150: 276–277

Flatz G (1987) Genetics of lactose digestion in humans. Adv Hum Genet 16:1–77

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–372

Hoegenauer C, Hammer HF, Mellitzer K, Renner W, Toplak H (2003) Evaluation of a new genetic test compared to the lactose hydrogen breath test for the diagnosis of acquired primary lactase deficiency. Gastroenterlogy Suppl 124:A64

Holden C, Mace R (1997) Phylogenetic analysis of the evolution of lactase digestion in adults. Hum Biol 69:605–628

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM (2001) Lactase haplotype diversity in the Old World. Am J Hum Genet 68: 160–172

Howell JN, Schockenhoff T, Flatz G (1981) Population screening for the human adult lactase phenotypes with a multiple breath version of the breath hydrogen test. Hum Genet 57: 276–278

Kayser M, de Knijff P, Dieltjes P, Krawczak M, Nagy M, Zerjal T, Pandya A, Tyler-Smith C, Roewer L (1997) Applications of microsatellite-based Y chromosome haplotyping. Electrophoresis 18:1602–1607

Kretchmer N, Ransome-Kuti O, Hurwitz R, Dungy C, Alakija W (1971) Intestinal absorption of lactose in Nigerian ethnic groups. Lancet 2:392–395

Kuokkanen M, Enattah NS, Oksanen A, Savilahti E, Orpana A, Jarvela I (2003) Transcriptional regulation of the lactase-phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. Gut 52:647–652

Kurt I, Abou Ghoush M, Hasimi A, Serdar M, Kutluay T (2003) Comparison of indirect methods of lactose absorption. Turk J Med Sci 33:103–110

Mathias N, Bayes M, Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. Hum Mol Genet 3:115–123

Murdock G (1967) Ethnographic atlas. University of Pittsburgh Press, Pittsburgh

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Newcomer A, McGill DB, Thomas P, Hofmann A (1975) Prospective comparison of indirect methods for detecting lactase deficiency. N Engl J Med 24:1232–1235

Newman J (1995) The peopling of Africa: a geographic interpretation. Yale University Press, New Haven, CT

Nurse G, Jenkins T (1974) Lactose intolerance in San populations. Br Med J 2:728

O'Keefe S, Adam J (1983) Primary lactose intolerance in Zulu adults. S Afr Med J 63:778–780

O'Keefe S, Adam J, Cakata E, Epstein S (1984) Nutritional support of malnourished lactose intolerant African patients. Gut 25:942–947

Olatunbosun D, Adadevoh B (1971) Lactase deficiency in Nigerians. Am J Dig Dis 16:909–914

Olds LC, Sibley E (2003) Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a *cis* regulatory element. Hum Mol Genet 12:2333–2340

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63: 1839–1851

Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarner M, Korpela R, Swallow DM (2003) The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. Ann Hum Genet 67:298–311

Puehkuri K (2000) Lactose, lactase and bowel disorders. PhD thesis, Hakapaino, Helsinki, http://ethesis.helsinki.fi/julkaisut/laa/biola/vk/peuhkuri/ (accessed April 5, 2004)

Ransome-Kuti O, Kretchmer N, Johnson J, Gribble J (1972) Family studies of lactose intolerance in Nigerian ethnic groups. Pediatr Res 6:359

——— (1975) A genetic study of lactose digestion in Nigerian families. Gastroenterology 68:431–436

Sahi T (1974) The inheritance of selective adult-type lactose malabsorption. Scand J Gastroenterol Suppl:1–73

Segal I, Gagjee P, Essop A, Noormohamed A (1983) Lactase deficiency in the South African black population. Am J Clin Nutr 38:901–905

Swallow DM, Harvey CB (1993) Genetics of adult-type hypolactasia. Dyn Nutr Res 3:1–7

Swallow DM, Hollox EJ (2000) The genetic polymorphism of intestinal lactase activity in adult humans. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic and molecular basis of inherited disease, 8th ed. McGraw-Hill, New York

Thomas MG, Bradman N, Flinn HM (1999) High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. Hum Genet 105:577–581

Troelsen JT, Olsen J, Moller J, Sjostrom H (2003) An upstream polymorphism associated with lactase persistence has increased enhacer activity. Gastroenterology 125:1686–1694

Y-chromosome consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 12:339–348